Response to Reviewer

Testing the Temporal and Spatial Transferability of a Water Balance Model

Degenne et al.

We thank the reviewer for their careful reading of the manuscript and for their constructive and open-minded comments. We are pleased that the reviewer found the idea 'simple and elegant' and identified no major flaws in the methodology. Below, we address each comment in turn.

General Comments

GC1 — Reference model: why a linear regression?

The reference approach is too simple in my view, and a non-linear one could be used as others use them in regionalization [...] why a comparison is made using a linear model as a reference?

We would like to clarify the rationale behind our choice of the local linear regression as a reference, which we believe may not have been sufficiently explained in the manuscript.

The so-called local linear regression benchmark is not intended to represent the state of the art in regionalization. It is important to distinguish two things here: the model structure and the parameter estimation method. The annual anomaly model at the core of our framework has four parameters: three elasticity coefficients (α_P, α_E, α_Λ) and one correcting factor for the long-term mean runoff (MQ). In what we call the hybrid approach, these four parameters are estimated by a neural network (NN) trained globally across all catchments using catchment descriptors — this is the differentiable parameter learning (dPL) strategy. In the local regression approach, the same four parameters of the same model are instead estimated independently for each catchment by ordinary least squares, using only local data and no catchment descriptors. Because this local fit uses all available data for a given catchment, it yields the best possible parameter set that our model structure can achieve for that basin; this is the LOCAL FIT configuration.

The comparison with LOCAL FIT is therefore a comparison with a clear reference: it answers the question 'can a globally-trained neural network using catchment descriptors be close to the best local calibration one could ever achieve for each individual basin?' A non-linear regionalization benchmark (e.g., MPR/mHM-style) would be comparing our approach against a different modelling system and a different parameter estimation strategy; we deliberately chose not to include such comparisons in order to focus specifically on the contribution of the neural network as a parameter estimation method within our own model structure, whereas here we isolate the contribution of the neural network relative to the best achievable local fit of our own structure.

Also, we would like to mention that our annual model is not purely linear. The Turc–Mezentsev formulation used to estimate MQ is a nonlinear function of precipitation and potential evapotranspiration. The linearity only applies to the anomaly component (i.e., how annual deviations from the long-term mean respond to climate anomalies). The framework therefore already incorporates a physically-grounded nonlinearity for the mean state.

We will revise the manuscript to make this justification explicit in the benchmark description section, as we recognise it is a key point for interpreting the results.

GC2 — Catchment IDs and data availability

No mention of the selected catchment IDs used in this study [...] You can also share your final training and validation datasets along with the code, if possible.

We agree that disclosing the list of catchment IDs is essential for reproducibility. We will provide the complete list of the 3,044 catchment IDs used in this study as supplementary material.

Regarding the input data, all CAMELS datasets used in this work are publicly available from their respective repositories, as referenced in Tables 1, A1, A2, and A3 of the manuscript. All catchment descriptors used to train the neural network are also drawn from these openly accessible sources. We will add an explicit statement to the Code and Data Availability section to make this clearer.

Regarding the code, sharing a fully reproducible version is more complex given the specific computational environment used (Jean-Zay supercomputer). The reviewer's interest in the code is noted; we acknowledge that full code sharing would require significant additional work to document and adapt the implementation beyond the specific computational environment used (Jean-Zay supercomputer), and we cannot commit to this at this stage.

Specific Comments

SC1 — L13: Gradients propagated through the whole modeling chain

What gradients propagated through the whole modeling chain?

We will clarify this sentence. The gradients referred to are the gradients of the loss function (MSE on streamflow) with respect to the neural network weights. Because the anomaly model equations are fully differentiable with respect to their parameters, these gradients are backpropagated from the streamflow output through the anomaly model equations and into the neural network weights, using automatic differentiation (PyTorch). We note that MQ, estimated via the Turc–Mezentsev equation, is pre-computed prior to training and acts as a fixed input to the anomaly model; it is therefore not part of the differentiable chain. We will rewrite this sentence to make it self-explanatory.

SC2 — L264-266 / L511: Choice of MSE as loss function

Is it correct to use MSE? Won't locations with an overall lesser precipitation/flow be disadvantaged? [...] Using NSE with MSE does not make sense.

We thank the reviewer for raising this point. The choice of MSE over NSE as a training criterion is motivated by the structure of our optimization problem. NSE is defined at the catchment level and requires a temporal series of observations to compute the mean flow used as its denominator — it is inherently a metric designed for models that produce continuous temporal

streamflow sequences. In our framework, however, the annual model computes each year independently, without any temporal link between successive years: it does not produce a streamflow time series but rather a collection of independent annual values. As a result, NSE is not a natural training criterion here, since there is no temporal continuity to exploit. In addition, the loss function is computed globally across all years and all catchments simultaneously, without any explicit grouping by basin, further reinforcing the appropriateness of MSE as an optimization target. We acknowledge, however, that MSE may give more weight to catchments with higher mean flows, as larger absolute residuals will dominate the loss. Alternative approaches — such as variance-based weighting across catchments or flow transformations prior to computing the loss — could mitigate this effect and would be worth exploring in future studies to assess the sensitivity of the results to the choice of training criterion.

The use of NSE as an evaluation metric (not as objective function) remains entirely consistent with this: NSE is computed post-hoc for each catchment once all annual predictions are available, and serves as a normalized diagnostic tool to compare performance across basins with different mean flows. We will clarify this distinction explicitly in the manuscript.

SC3 — L374-378: Parameter ranges not hard-coded

How does the HYBRID approach know what is physical when nobody told it? […] give it random combinations of inputs […] check if they still lie between 0 and 1.

We confirm that no constraint is hard-coded in the model to force the parameters into physically plausible ranges. There is no sigmoid, clipping, or bounding operation applied to the neural network outputs. The fact that the parameters spontaneously fall within plausible ranges is an emergent property of training on real catchment data with a physically structured model.

The proposed test, i.e. feeding random input combinations and checking whether outputs remain in [0, 1], is an interesting diagnostic. We note that such a test would evaluate extrapolation behaviour beyond the data manifold, which is a genuinely open question. We will mention this as a suggested avenue for future robustness testing in the discussion.

SC4 — L279-285: Hyperparameter stability across runs

Were the 77 combinations also done multiple times and did each time, the 8 hidden layers and 192 units was the best solution?

Yes. Each of the 77 configurations was trained multiple times with different random seeds. The grid search revealed that beyond a certain level of network complexity, additional layers or units yielded only marginal improvements in validation performance, i.e. the performance curve reaches a plateau. Several architectures in this plateau region produced similar results. We selected the configuration with 8 hidden layers of 192 units each as it was the most parsimonious architecture that had already reached this plateau, in line with the parsimony principle advocated throughout the study. This choice was therefore not arbitrary but motivated by the desire to avoid unnecessary model complexity once the performance plateau had been reached.

Each individual training run takes only a few minutes on the Jean-Zay GPU infrastructure (Nvidia V100, 32 GB), making the exhaustive grid search of 77 configurations computationally feasible. We will add both the computation time and the plateau observation in the revised manuscript.

SC6 — L341-350: Statistically insignificant regression coefficients

Maybe, a bit more information can be added here.

We will clarify that a regression coefficient is considered statistically insignificant when its associated p-value exceeds 0.05, as computed by the OLS routine of the statsmodels Python library (two-sided t-test). In such cases, the coefficient is set to zero, meaning the corresponding predictor is excluded from the local regression model for that catchment.

SC7 — L356-364: Calibration before validation in results

Maybe discuss the calibration before the validation?

We agree and will reorganize the presentation of results to follow the more natural order: calibration setting (LOCAL FIT vs. HYBRID FIT) first, then temporal cross-validation (LOCAL TEMP vs. HYBRID TEMP). This will allow readers to first understand the upper bound of performance before seeing how each approach degrades under cross-validation.

SC8 — L250-255: Number of trainable parameters

Any idea on how many parameters need to be calibrated for this dense neural network?

The total number of trainable parameters in the neural network can be computed from the selected architecture (8 hidden layers of 192 units each, input layer of 27 descriptors, output layer of 4 parameters). Accounting for weights and biases across all layers, the network contains approximately 260,000 trainable parameters. This is substantially larger than a traditional regionalization model, but the network is trained on a dataset of 3,044 catchments covering several decades of annual data. We acknowledge that the total number of observations (catchments × years) may in fact be of the same order of magnitude as the number of trainable parameters, which raises a legitimate question about the degree of over-parameterisation. This is a known challenge in deep learning applied to hydrology, and it is partly mitigated by the strong implicit regularisation brought by the physically-structured model backbone and the shared learning across all catchments. Whether a more parsimonious network architecture would achieve comparable or better regionalisation performance is an open and interesting question that we will flag in the revised manuscript as a perspective for future work. We will add this count explicitly to the manuscript.

SC9 — L275-279: Variability across spatial splits

Is it shown how different the results of these five models are? Very different results in general would mean that the model is set up, possibly, incorrect.

This is a relevant diagnostic. In the spatial cross-validation, 10 training configurations were tested (corresponding to all combinations of 3 groups among 5). The variability in NSE_BOUNDED across these 10 configurations was relatively limited, indicating that the results

are stable with respect to the choice of training basins. We did not present this variability explicitly in the manuscript, but we will add a brief note; for instance as a range or standard deviation across configurations; to confirm the robustness of the results and address the reviewer's concern.

SC10 — L335: Small sample sizes after data thinning

The table shows extreme thinning of the data, from what I understand. I wonder how much can be learned from such small samples.

The reviewer raises a valid concern. The thinning visible in the table results from the requirement that catchments have at least 6 annual values in each of the three temporal periods, a strict but necessary quality filter, to ensure meaningful cross-validation. We acknowledge that some catchments contribute relatively few data points. However, the key strength of the differentiable parameter learning (dPL) framework is that learning is performed globally across all catchments simultaneously: the neural network does not learn from each catchment in isolation, but from the collective signal of the full 3,044-catchment dataset. Even catchments with short records contribute to constraining the shared parameter-descriptor relationships. The small per-catchment sample size is therefore less limiting than it would be in a local calibration context. We will clarify this point in the revised manuscript.

SC --- L58: Model parsimony vs. neural network

Here model parsimony is mentioned but I am curious to see how, later on, a model i.e., neural network, which goes against this philosophy, is selected and evaluated.

This is a fair observation. We use parsimony as a guiding principle for the hydrological model structure (4 parameters, annual time step), not as an absolute rule applied to every component of the framework. The neural network is the parameter estimation method, not the hydrological model itself; its complexity is justified by the size of the multi-country dataset (3,044 catchments) and the need to learn complex, nonlinear relationships between catchment descriptors and model parameters. We acknowledge this apparent tension and will address it explicitly in the revised manuscript.

SC --- L161-167

A good idea.

Thank you.

SC --- L189-191: Spatial averaging of gridded climate data

Taking the average of the gridded values is more correct, but I think the difference is not large from using the whole catchment temperature.

We agree with the reviewer's assessment. The spatial averaging of gridded values is the approach applied in our study, and we concur that the difference with a catchment-wide average is unlikely to be significant at the annual time step.

SC --- L195: List of catchment IDs

Is there a list of all the catchment IDs somewhere?

This is addressed in our response to GC2 above: we will provide the complete list of the 3,044 catchment IDs as supplementary material in the revised manuscript.

SC --- L247-249: Hundecha and Bardossy (2004)

Now that I think about it, this is what Hundecha and Bardossy (I think in 2004) were doing when they were regionalizing HBV parameters somewhere in Germany. The local-calibration-first and later-fitting has been obsolete since then, IMO. Some people still do it.

Thank you for this historical perspective. We agree that end-to-end approaches that avoid the two-step local-calibration-then-fitting procedure represent the current direction of the field, which is precisely the paradigm adopted in our dPL framework.

SC --- L295: Discarding catchments

Let's hope that the discarding does not lead to surprises later.

We share this concern. The catchments discarded are those with fewer than 6 annual values per period, which we judged insufficient for meaningful cross-validation. We do not expect systematic biases from this filtering, as the criterion is purely data-quantity based rather than performance-based. However, we acknowledge that this may introduce a slight selection bias toward catchments with longer and more complete records, which we will note in the manuscript.

SC --- L299-301: Temporal period in spatial validation

I think, it is implicitly meant that for spatial validation, time period is fixed. Maybe, mention it just to not have any confusion.

We would like to clarify that the temporal period is in fact not fixed in the spatial cross-validation experiment (HYBRID REGIO): all available time steps are used for each catchment regardless of the spatial split. The experiment that simultaneously varies both the catchment split and the time period is the spatiotemporal cross-validation (HYBRID SPATIO-TEMP). We will add a sentence to make this distinction explicit in the manuscript.

SC --- L372-374: Non-linear reference model

Well, linear regression is used and it is most probably that the regression is supposed to be non-linear. What if a non-linear regression was used as the reference? Linear relationships are only one possibility.

This point is addressed in detail in our response to GC1 above. In summary, the local linear regression is the optimal local calibration of our own model structure — the linearity is a property

of the anomaly component of the model, not an arbitrary choice of benchmark. Comparing against a non-linear external benchmark would conflate the model structure with the parameter estimation strategy.

SC --- L386-390 and L396-405: Spatial performance and anomaly approach

Interesting. Traditional techniques also fail at the same [...] All this time, we could have switched to anomaly [...] One could test the same for a conceptual daily rainfall-runoff model.

Thank you for these stimulating observations. The connection between our anomaly-based decomposition and the limitations of traditional approaches is indeed at the heart of the paper's contribution. The suggestion to test an anomaly framework on a conceptual daily model is an interesting avenue that we will mention as a perspective in the revised manuscript.

SC --- L519-521: Practical use case

A crucial use case. I agree.

Thank you for this comment. We are glad the reviewer shares our view on the practical relevance of this use case.

We hope these responses address the reviewer's concerns. We remain open to further discussion on any of these points.