

Replies to reviews

“Same Streamflow, Different Water Stories: The Hidden Impacts of Streamflow-Only Calibration in Distributed Hydrological Modeling”

Nicolás A. Vásquez, Pablo A. Mendoza, Wouter Knoben, Martyn Clark, Tricia Stadnyk, and Naoki Mizukami

We thank the reviewer for the time and constructive comments on our paper. We provide responses to each individual point below. For clarity, the reviewer's comments are given in italics, and our responses are given in plain blue text.

Reviewer #1

This manuscript addresses a highly relevant and persistent problem in distributed hydrological modeling: the extent to which streamflow-only calibration compromises the fidelity of simulated internal states and fluxes. By systematically exploring combinations of objective functions and spatial regularization strategies across six diverse Chilean catchments, the authors provide compelling evidence that similar streamflow performance can emerge from substantially different internal hydrological dynamics. The manuscript is generally well written, the hypotheses are clearly stated. However, several technical and presentation issues need to be addressed before publication.

We thank the reviewer for the positive assessment of our work and for the constructive comments. We address each point in detail in our point-by-point responses to the specific comments.

Specific comments:

1. The manuscript proposes a PCA-based spatial regularization strategy using physiographic variables. While this approach is interesting, the novelty relative to existing regionalization and transfer-function approaches (e.g., MPR) remains somewhat unclear. In addition, the approach appears somewhat empirical, the motivation for using only PC1 is insufficiently explained. Could the use of additional PCs, or a weighted combination of PCs, improve the transferability of the approach? The physical meaning of the PCA-derived spatial fields is insufficiently discussed.

We thank the reviewer for raising this important point. We agree that the conceptual relationship between our PCA-based spatial regularization strategy and existing regionalization/transfer-function approaches, such as Multiscale Parameter Regionalization (Samaniego et al., 2010; Kumar et al., 2013), was not sufficiently articulated. In the revised manuscript, we will add the following text in Section 4.2 to contrast the two approaches:

“Note that the regularization approach implemented in this study is very similar to the one proposed by Pokhrel et al. (2008), except that we use PC1 extracted from several attributes instead of the spatial distribution of a single physiographic property as a predictor. On the other hand, our approach differs from the Multiscale Parameter Regionalization (MPR; Samaniego et al., 2010) method in that the latter (i) directly links physiographic properties to model parameters through prescribed linear or nonlinear transfer functions, and (ii) considers upscaling operators to obtain model parameter values for the simulation scale.”

We only used the first component (PC1) because (i) it explained between 46% (Colorado River basin) and 79% (Cautín River basin) of the total variance in the physiographic variables considered across the six catchments, and (ii) PC1 was highly and significantly correlated with at least four grid cell attributes at Claro, Cautín and Futa, and two attributes in the remaining basins. In particular, correlations tend to be generally higher between PC1 and elevation, clay content and sand content

across basins. We have explicitly added this explanation in section 4.2, and we will update Figure S1 (Supporting Information) accordingly.

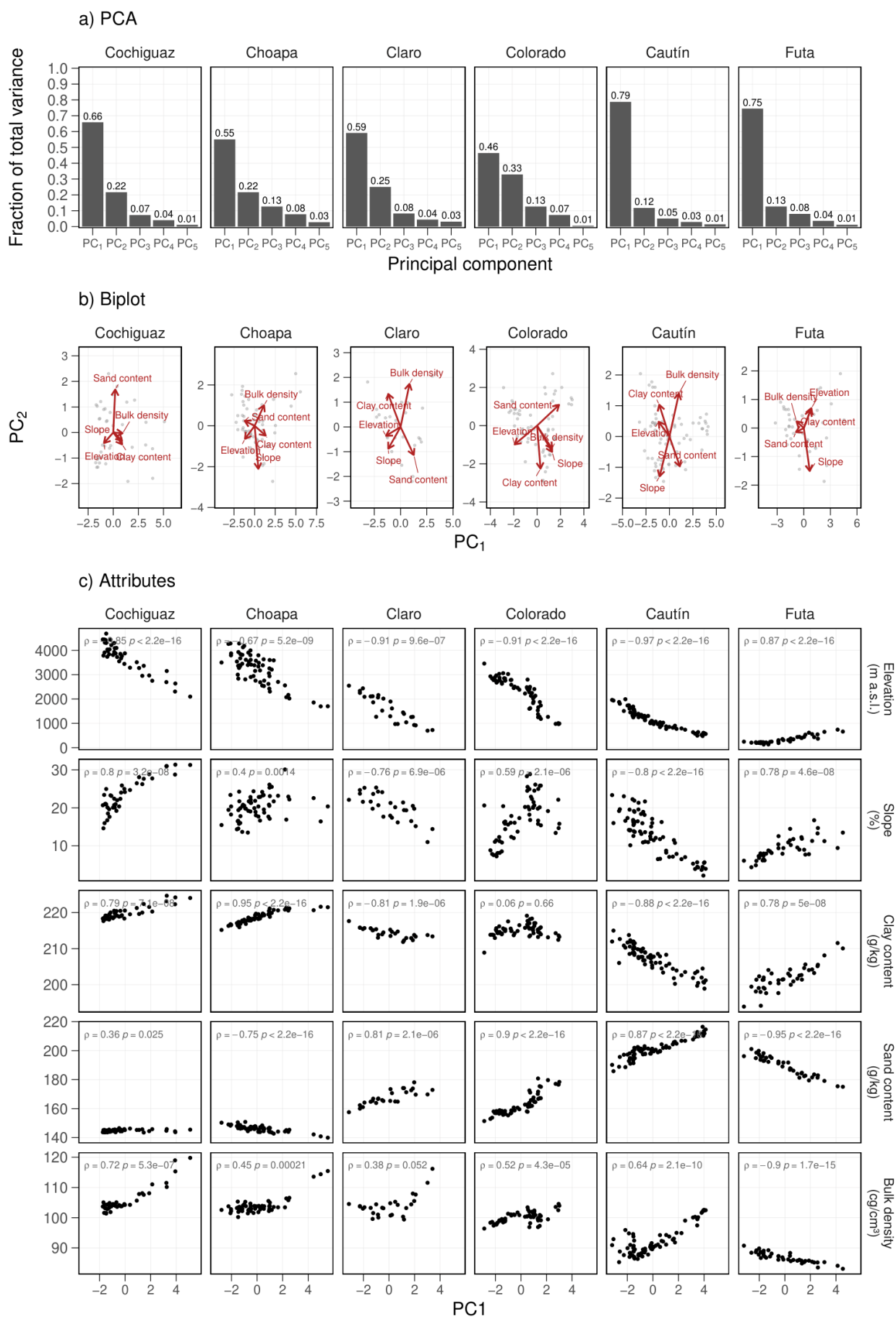


Figure S1. Principal component analysis to derive *a priori* parameter fields: a) fraction of total variance explained by each component; b) two first principal components biplot, and c) relationship between PC₁ and the attributes, along with Spearman rank correlation coefficients and p-values. Each dot represents a grid cell (0.05° x 0.05° horizontal resolution).

Regarding the number of PCs considered for spatial regularization, we conducted additional calibration experiments by maximizing KGE(Q) in order to compare the following cases: (i) spatially uniform parameter sets; (ii) distributing only b_{infiltr} as a function of PC₁ (i.e., $b_{\text{infiltr}} = a + b \cdot \text{PC}_1^c$); and (iii) distributing only b_{infiltr} as a function of PC₁ and PC₂ (i.e., $b_{\text{infiltr}} = a + b \cdot \text{PC}_1^c + d \cdot \text{PC}_2^c$). A comparison of simulated annual cycles of streamflow (Figure R1) and other variables (Figure R3), as well as daily flow duration curves (Figure R2), will be included in the Supporting Information section. One can note that very similar KGE(Q) values and streamflow annual cycles are obtained during the calibration period (Figure R1), and that the two regularization strategies applied to b_{infiltr} improve KGE(Q) during the evaluation period in Cochiguaz and Choapa. Additionally, the simulation of low flow volumes improves in Colorado and Futa when both PC₁ and PC₂ are considered for spatial regularization (Figure R2). However, the examination of simulated annual cycles of other variables than streamflow (Figure R3) reinforces that good streamflow simulations do not necessarily translate into annual cycles closer to observational references of other variables (see ET' results for Claro, Colorado, Cautín and Futa in Figure R3). We will include this point – i.e., that the main message of the manuscript does not change by including PC₂ in the parameter regularization scheme – in the discussion section of the revised manuscript.

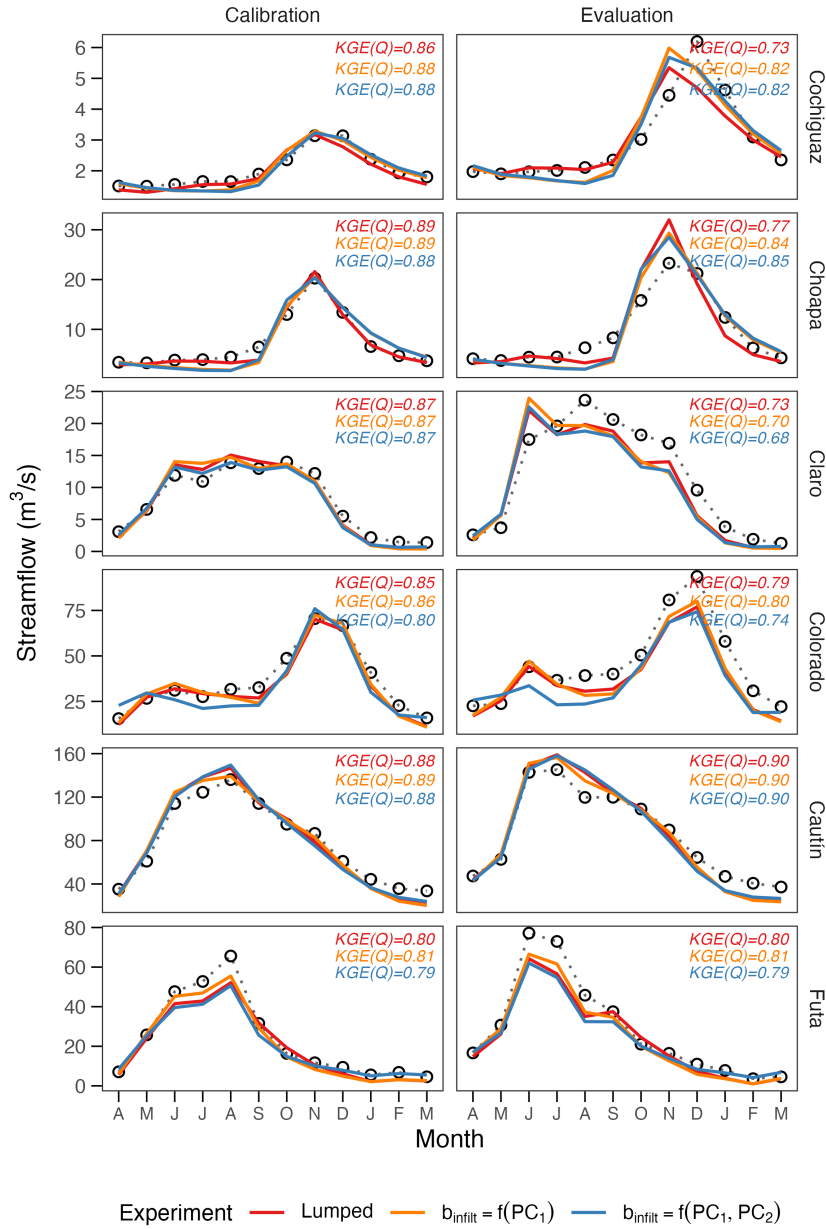


Figure R1. Simulated streamflow annual cycles for parameter sets obtained by maximizing KGE(Q) for the following cases: (i) spatially uniform parameter sets; (ii) distributing only b_{infil} as a function of PC_1 (i.e., $b_{\text{infil}} = a + b \cdot \text{PC}_1^c$); and (iii) distributing only b_{infil} as a function of PC_1 and PC_2 (i.e., $b_{\text{infil}} = a + b \cdot \text{PC}_1^c + d \cdot \text{PC}_2^e$). The calibration and evaluation periods are Apr/2005-Mar/2018 and Apr/1995-Mar/2005, respectively. The circles represent the observed mean monthly values.

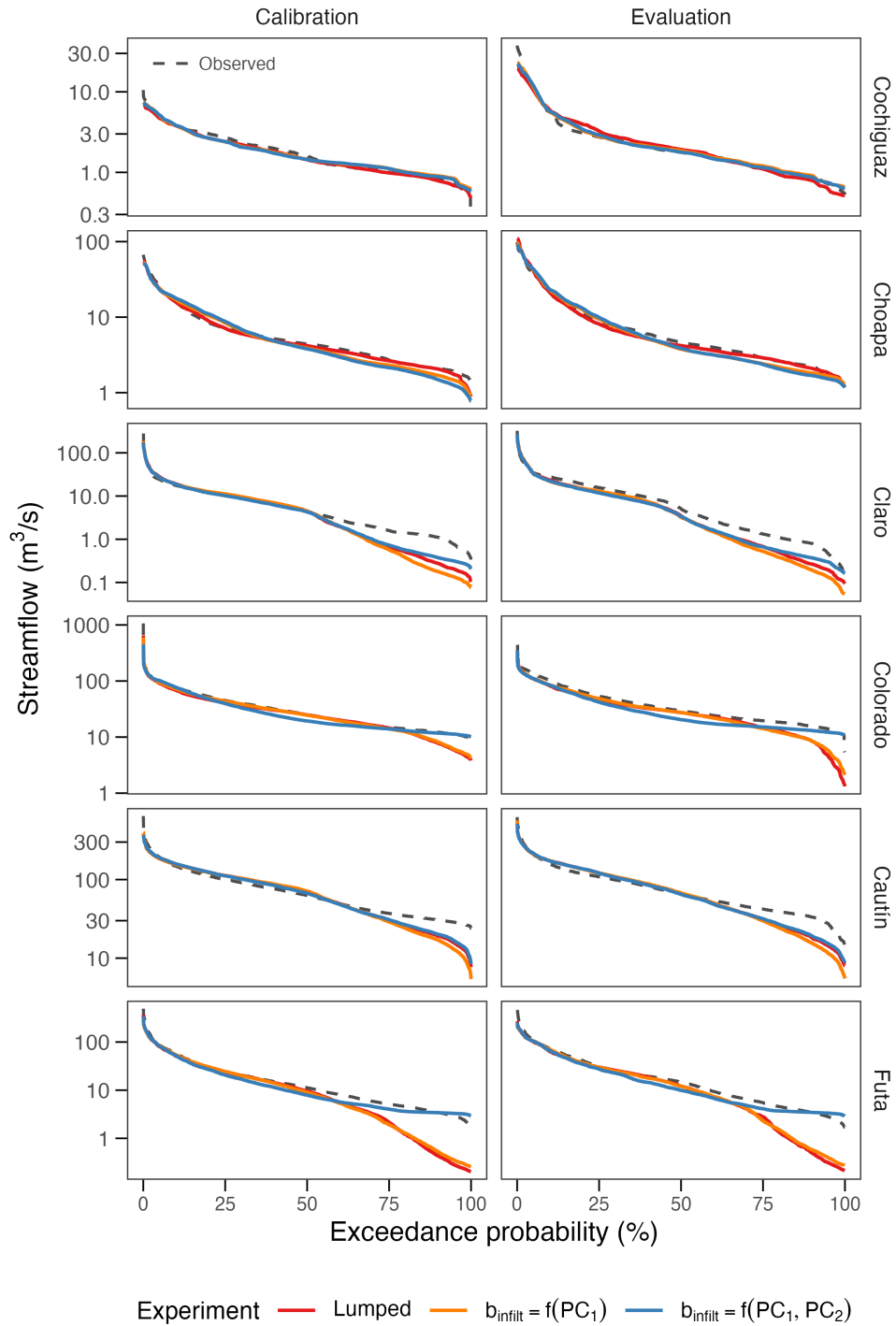


Figure R2. Simulated daily flow duration curves for parameter sets obtained by maximizing KGE(Q) for the following cases: (i) spatially uniform parameter sets; (ii) distributing only b_{infiltr} as a function of PC_1 (i.e., $b_{\text{infiltr}} = a + b \cdot \text{PC}_1^c$); and (iii) distributing only b_{infiltr} as a function of PC_1 and PC_2 (i.e., $b_{\text{infiltr}} = a + b \cdot \text{PC}_1^c + d \cdot \text{PC}_2^e$). The calibration and evaluation periods are Apr/2005-Mar/2018 and Apr/1995-Mar/2005, respectively.

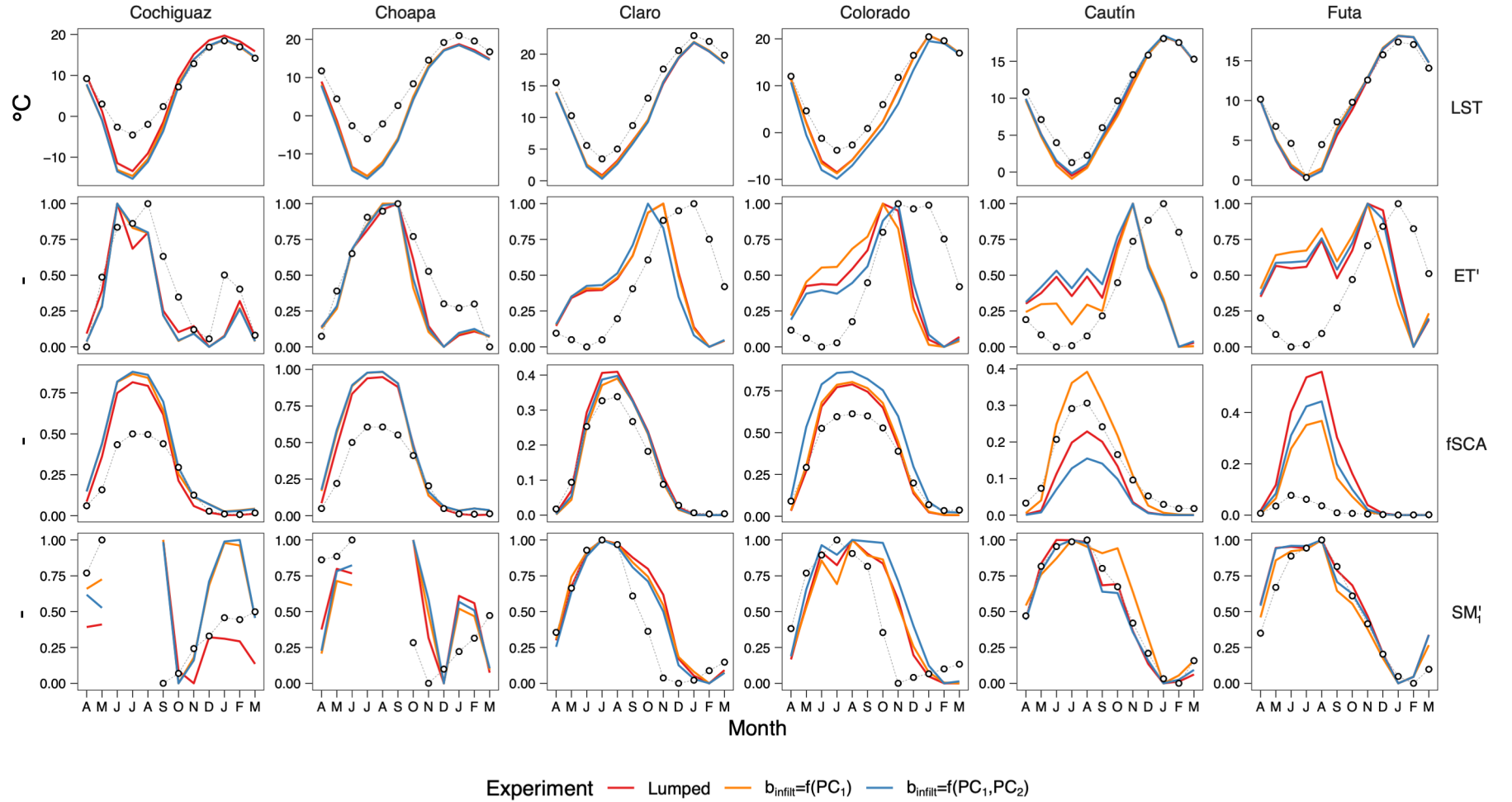


Figure R3. Simulated annual cycles of Land Surface Temperature (LST), dimensionless Evapotranspiration (ET'), fractional Snow Covered Area (fSCA) and Soil Moisture for the first (upper) soil layer (SM_{1'}) for parameter sets obtained by maximizing KGE(Q) for the following cases: (i) spatially uniform parameter sets; (ii) distributing only b_{infil} as a function of PC₁ (i.e., $b_{infil}=a+b*PC_1^c$); and (iii) distributing only b_{infil} as a function of PC₁ and PC₂ (i.e., $b_{infil}=a+b*PC_1^c+d*PC_2^e$). The circles represent the mean monthly observational reference values, and results are displayed for the calibration period.

2. The manuscript repeatedly emphasizes equifinality and compensation among fluxes and states. However, the study does not provide a formal parameter uncertainty or identifiability analysis. I'm wondering how different objective functions influence parameter constraints, whether the inclusion of ET or SM observations reduces parameter equifinality, and to what extent the proposed PCA-based regularization framework effectively improves parameter identifiability and mitigates equifinality in distributed hydrological modeling.

We thank the reviewer for this comment. We would like to stress, however, that this study does not aim to provide formal parameter uncertainty and/or parameter identifiability analyses; instead, we seek to explore the extent to which the choice of Q-only objective functions and spatial regularization of parameters alter other variables (besides streamflow) that are relevant to characterizing catchment-scale seasonal water balances. Additionally, a key contribution of this study is the examination of flux equifinality – which is only one facet of hydrological model equifinality (Khatami et al., 2019) – in average annual cycles. We will clarify this point at the end of the introduction of the revised manuscript. Further, we will clarify in the discussion section that spatial parameter regularization strategies are not intended to improve parameter identifiability and/or mitigate equifinality; instead, they seek to reduce the dimensionality of the parameter estimation problem in distributed hydrological modeling (Pokhrel et al., 2008; Pokhrel & Gupta, 2010; de Lavenne et al., 2019).

Nevertheless, we did verify the normalized parameter values associated with different regularized parameters (Figure R4) obtaining that, as expected, very different parameter sets can lead to similar streamflow performance metrics. We also conducted exploratory analyses (section S.6 in the Supporting Information section) to examine – for a given Q-based calibration metric – the interplay between equifinal parameter sets and the regularization scheme for simulating annual cycles of Q, ET' and SM', LST and fSCA (Figure S34, Figure R5 in this document). The results show that ET' seasonality can vary substantially among near-optimal parameter sets depending on the combination of regularization scheme and catchment (e.g., spatially constant parameters at Cautín; regularized Depth₂ at Colorado). These results suggest that modelers prioritizing internal process fidelity should verify the annual cycle of ET (and other relevant variables) using a small ensemble of behavioral parameter sets rather than the single best-fit solution. We have incorporated these points at the end of section 6.2 (Discussion).

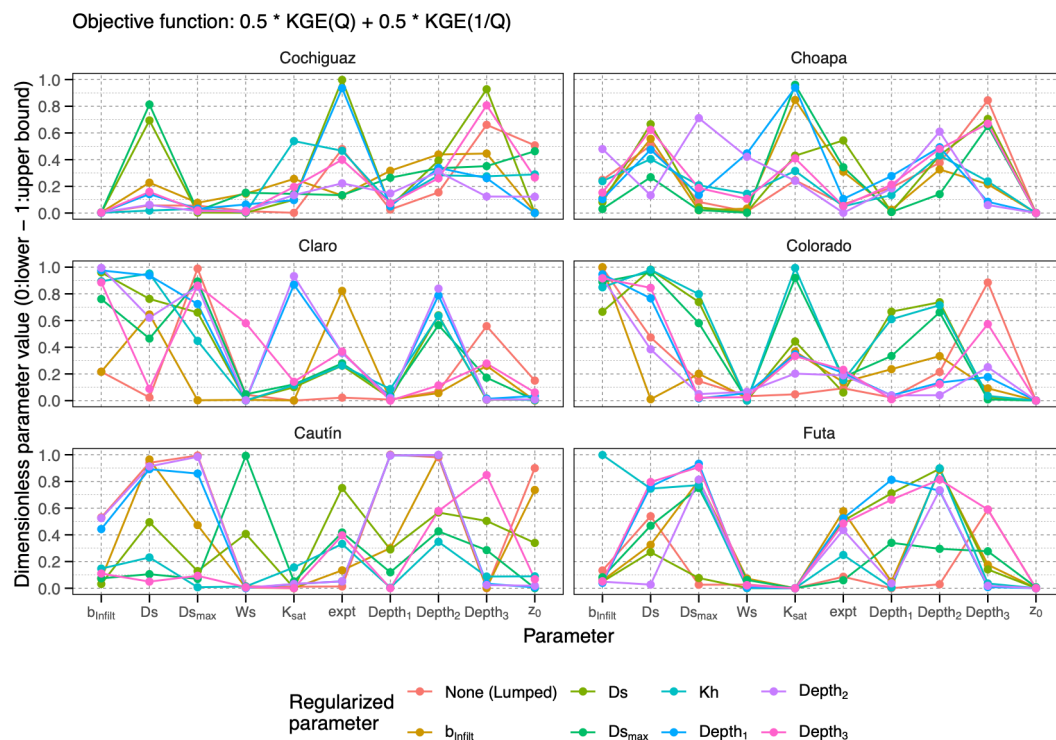


Figure R4: Normalized Parameter values – computed as $X' = (X - X_{\min}) / (X_{\max} - X_{\min})$ – obtained from the calibration process with all configurations, using $0.5 \cdot KGE(Q) + 0.5 \cdot KGE(1/Q)$ as the objective function. For regularized parameters, we show the basin average.

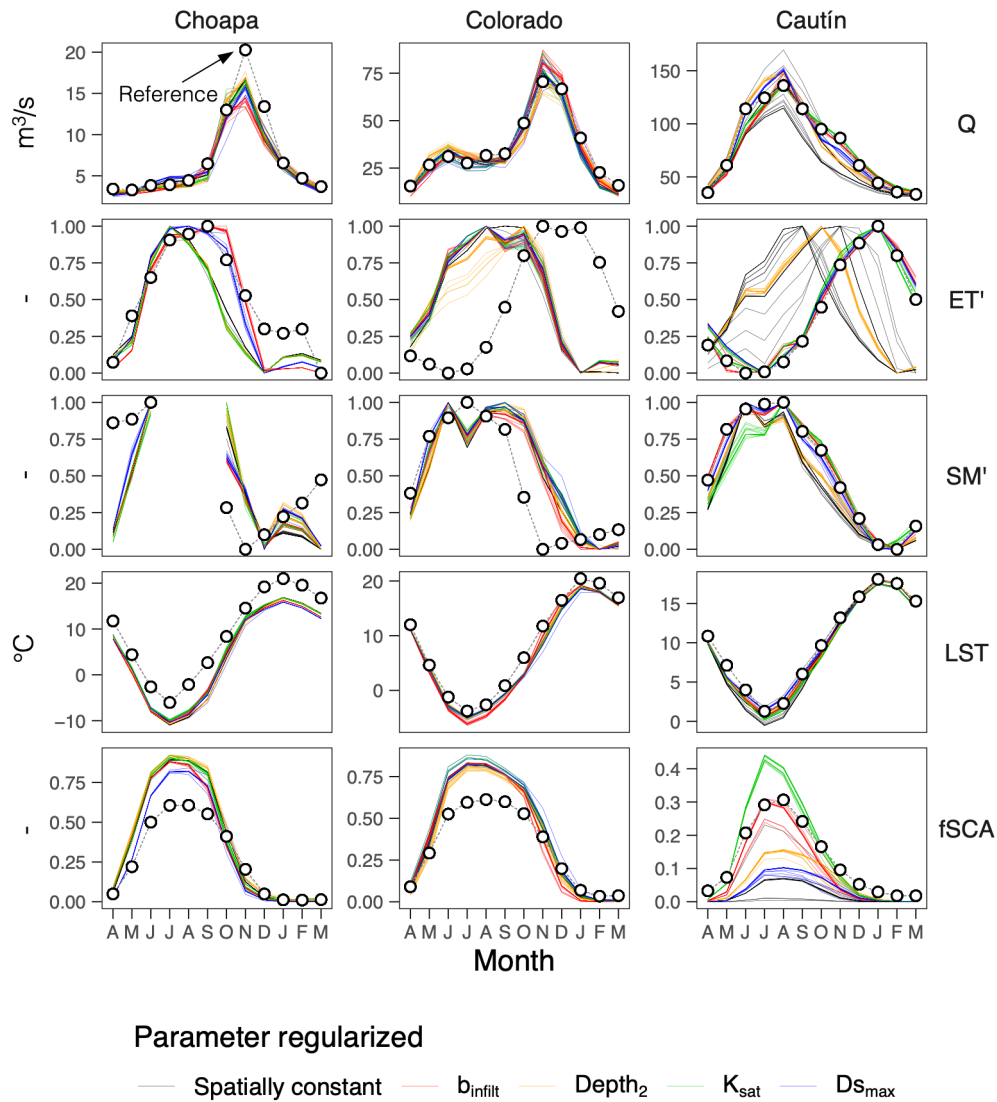


Figure R5. Catchment-scale annual cycles of streamflow (Q), normalized evapotranspiration (ET'), normalized soil moisture (SM'), land surface temperature (LST) and fractional snow covered area (fSCA) for the Choapa, Colorado, and Cautín River basins during the calibration period (2005–2018). Each line corresponds to a behavioral parameter set selected using two criteria: (1) the best 1% of parameter sets ranked by objective function value, or (2) all parameter sets with an objective function value exceeding $\max[OF(Q)] - 0.02$, where $OF = 0.5 \cdot KGE(Q) + 0.5 \cdot KGE(1/Q)$. Streamflow observations and remotely sensed variables are referred to as “reference” and represented with white dots. The normalization of ET and SM₁ is computed as $X' = (X - X_{\min}) / (X_{\max} - X_{\min})$. Note that winter corresponds to JJA, while summer to DJF.

Additionally, we examined the sensitivity of our results and conclusions to the initial parameter set and seed number used in DDS optimizations for the following cases: (i) spatially uniform parameter sets; (ii) distributing only b_{infil} as a function of PC_1 (i.e., $b_{\text{infil}} = a + b \cdot PC_1^c$); and (iii) distributing only b_{infil} as a function of PC_1 and PC_2 (i.e., $b_{\text{infil}} = a + b \cdot PC_1^c + d \cdot PC_2^e$). We performed these calibration experiments only at the Futa River basin (i.e., the smallest one), to reduce the computational burden. The results show that, despite better $KGE(Q)$ values being achievable with different random seeds and including both PC_1 and PC_2 in the regularization scheme (Figure R6), none of the new parameter sets yield reasonable annual cycles of variables other than streamflow (Figure R7). We have incorporated these points in section 6.2 (Discussion).

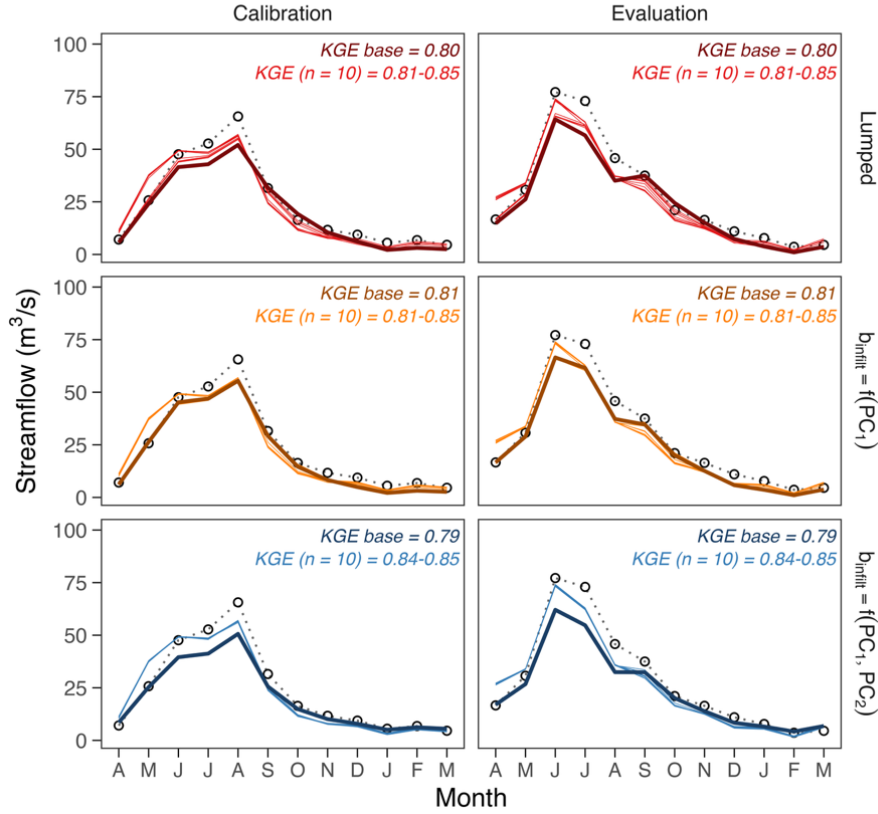


Figure R6: Streamflow annual cycles for additional calibrations conducted for the Futa River Basins to verify the impact of different initial parameter sets and seeds. In all the experiments, KGE(Q) was maximized using DDS (budget = 1,000 simulations), and results are presented for the following cases: (i) spatially uniform parameter sets; (ii) distributing only b_{infil} as a function of PC_1 (i.e., $b_{infil} = a + b \cdot PC_1^c$); and (iii) distributing only b_{infil} as a function of PC_1 and PC_2 (i.e., $b_{infil} = a + b \cdot PC_1^c + d \cdot PC_2^e$). The thicker lines represent the original results to facilitate comparisons, and the circles represent the mean monthly observational reference values.

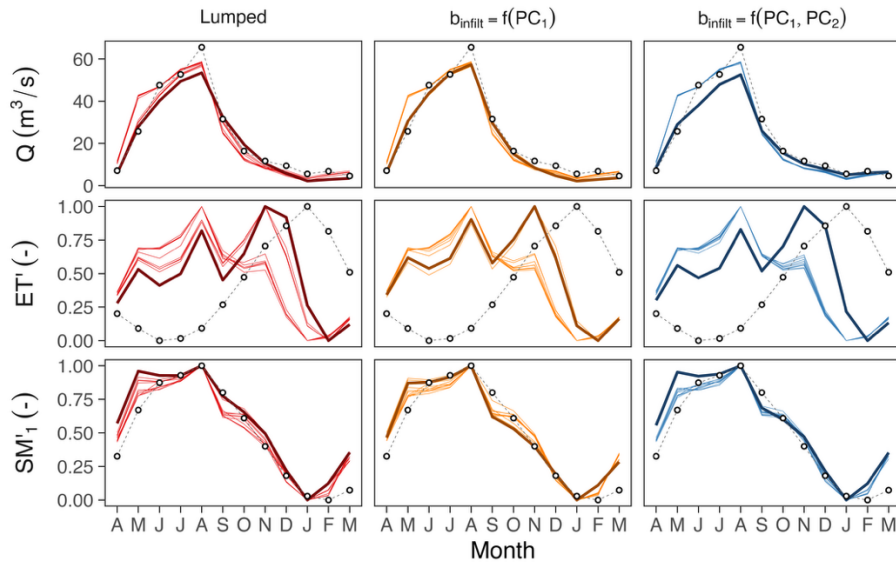


Figure R7: Simulated annual cycles of streamflow, dimensionless Evapotranspiration (ET') and dimensionless Soil Moisture for the first (upper) soil layer (SM'_1) for parameter sets obtained by maximizing KGE(Q) with DDS (budget = 1,000 simulations) and different initial parameter sets and seeds. The thicker lines represent the original results to facilitate comparisons, and the circles represent the mean monthly observational reference values.

3. One of the key findings is that ET seasonalities differ substantially despite similar streamflow simulations. This is an interesting and important result. The analysis relating these ET shifts to moisture availability in deeper soil layers (Figure 9) is insightful. However, the discussion stops short of a full diagnostic. It would be beneficial to explicitly state which process parameterizations are primarily impactful in the configurations that produce the most erroneous ET seasonalities. For instance, are large ET biases consistently linked to unrealistic soil water storage dynamics in layer 3? Whether snow accumulation/melt timing contributes to ET discrepancies? A more mechanistic interpretation would elevate the paper's impact.

We thank the reviewer for this suggestion. With our experimental design, it is not possible to modify flux parameterizations (also referred to as closure relationships) in the VIC model, since the model structure is fixed. However, we will extend Figure S34 to break simulated ET into its components and show them next to SWE, runoff, baseflow and surface soil moisture (SM_1). The new results (see preliminary version, Figure R8) show that a good match between simulated and reference ET' seasonalities can be achieved through strong flux compensation. In Choapa ($b_{infiltr}$ and Ds_{max}) and Cautín ($b_{infiltr}$, K_{sat} and Ds_{max}), several configurations match the reference ET' seasonality and streamflow equally well, with very different canopy evaporation, transpiration, SWE, surface runoff and baseflow. As a result, no single configuration can be identified as the source of the ET errors. We will incorporate this diagnosis in Section 5.3.

In this study, we used a single SRS ET product, which we use only as a normalized seasonal signal and does not separate transpiration from evaporation. Additionally, the soil moisture product covers only the top few centimeters (soil layer 1). In Cautín, all configurations match the reference surface soil moisture seasonality, including those with 'wrong' ET (Figures 5 and 6). Soil layer 1 therefore does not constrain the deeper storage that controls transpiration, and soil layers 2 and 3 have no observational constraint at all. We have included these points in Section 6.3 (Discussion).

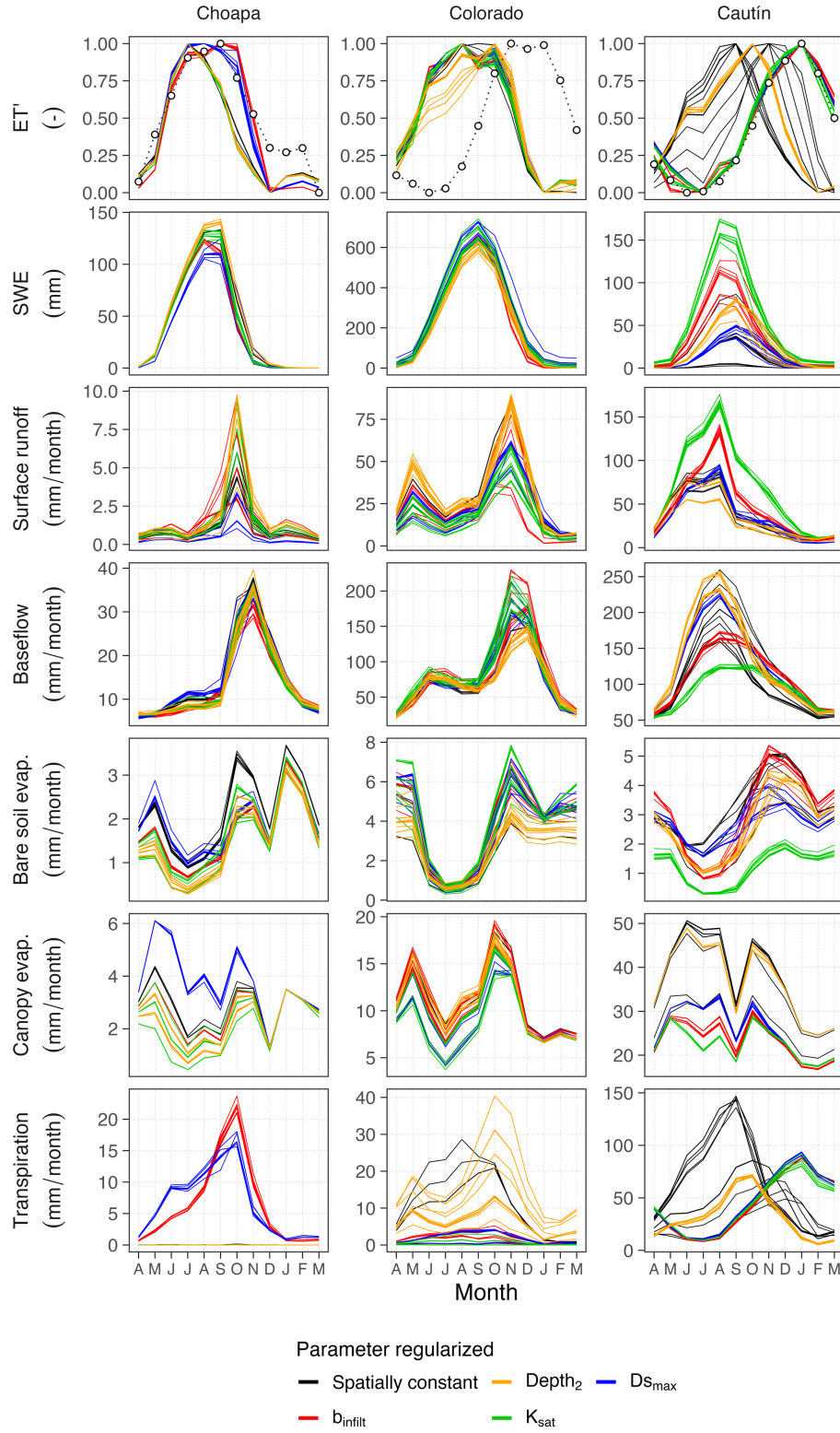


Figure R8. Catchment-scale annual cycles of normalized evapotranspiration (ET'), snow water equivalent (SWE), surface runoff, baseflow, bare soil evaporation, canopy evaporation and transpiration for the Choapa, Colorado, and Cautín River basins during the calibration period (2005–2018). Each line corresponds to a behavioral parameter set selected using two criteria: (1) the best 1% of parameter sets ranked by objective function value, or (2) all parameter sets with an objective function value exceeding $\max[\text{OF}(Q)] - 0.02$, where $\text{OF} = 0.5 \cdot \text{KGE}(Q) + 0.5 \cdot \text{KGE}(1/Q)$. Streamflow observations and remotely sensed variables are referred to as “reference” and represented with white dots. The normalization of ET and SM₁ is computed as $X' = (X - X_{\min}) / (X_{\max} - X_{\min})$. Note that winter corresponds to JJA, while summer to DJF.

4. The additional two multi-objective calibration experiments are valuable but appear somewhat preliminary relative to the broader conclusions of the paper. Why did the authors not consider combining multivariate calibration with the proposed spatial regularization framework?

We thank the reviewer for this comment. In this study, we decided to focus on the effects of Q-only calibration across different regularization configurations, since it is common practice in hydrological modelling. Hence, we only used Q-based calibration metrics, including two meta-objective functions (which condense two or more metrics, see equations 3 and 4) in our experiments to isolate the individual effects of each strategy before combining them. Although the multivariate (Q+SM and Q+ET) calibration experiments included in this study are not central, they complement the results and insights presented in this work, and we agree with this reviewer that further analyses on the impacts that multivariate calibration metrics and regularization strategies exert on simulated hydrological variables is the logical next step, which is part of ongoing research in our group. We clarify this point in the Discussion section:

“Although we conducted additional calibration experiments by adding more variables to the OF (Q-ET and Q-SM₁), we stress that these experiments are exploratory, since they did not consider the PCA-based spatial regularization strategy proposed here. Additionally, we did not conduct a systematic evaluation that includes all variables and basins, since the associated benefits have been widely discussed in the literature (López et al., 2017; Demirel et al., 2018; Koppa et al., 2019; Dembélé et al., 2020; Shah et al., 2021; Tong et al., 2021; Pool et al., 2024). A logical next step is the assessment of the impacts that combined multivariate calibration metrics and regularization strategies exert on simulated hydrological variables, which is part of ongoing research.”

Additionally, we will revise the framing of the two additional multi-objective experiments to present them explicitly as complementary and exploratory analyses rather than as central pillars of the paper's conclusions (Section 4.3.3):

“To complement the Q-only calibrations presented here and explore the potential effects of explicitly incorporating ET or SM₁ in the objective function, we conduct two additional calibration experiments using spatially constant parameters to minimize an Euclidean-based objective function:”

5. Although the manuscript explicitly states that one of the research questions is how to overcome the tradeoffs between accurately replicating streamflow annual cycles and effectively simulating the seasonal patterns of other hydrological variables, the study does not appear to provide a clear or systematic answer to this question beyond demonstrating the existence of such tradeoffs.

We really appreciate this observation. We believe that, in order to provide a robust strategy to overcome the trade-offs between accurately replicating streamflow annual cycles and effectively simulating the seasonal patterns of other hydrological variables, many more calibration experiments would be required, considering (1) uncertainties in the meteorological forcings, (2) uncertainties in the reference observational datasets (especially the SRS products), and (3) alternative hydrological models, given the possibility that the VIC model structure may not be flexible enough to compensate for other sources of errors through the parameter estimation process. Because of these reasons and in response to the reviewer's observation, we have decided to remove the second part of question 2 (How can these trade-offs be overcome?). Additionally, we will incorporate these points as limitations that challenge the resolution of tradeoffs between replicating streamflow annual cycles and achieve realistic simulations of seasonal patterns of other hydrological variables (section 6.3), and we will reinforce the main message of this paper: for water balance applications of any hydrological models, the evaluation of simulated annual cycles should be considered, including the variable(s) of interest (in this case ET). To do so, we have modified the last sentence of the abstract as follows:

“Overall, the results presented here (1) reinforce the benefits of including remotely sensed spatial patterns in the calibration of distributed hydrological models, (2) highlight the need to verify the

seasonality of other simulated variables besides streamflow, and (3) stress the need to include ET within the calibration process for seasonal water balance portrayals.”

We have also modified the fourth conclusion as follows:

“For water balance characterizations, evaluating simulated annual cycles of variables other than streamflow is required to detect model or data deficiencies. In particular, the analyses presented here stress the need to include ET within the calibration process for seasonal water balance portrayals.”

6. The study only used a single calibration period (2005–2018) without independent validation. Given the strong conclusions regarding model realism and process fidelity, an independent validation period is important to demonstrate robustness and avoid overfitting.

Originally, we considered Apr/2005-Mar/2018 and Apr/1995-Mar/2005 as the calibration and evaluation periods, respectively (excepting the Futa River basin, where calibration and evaluation periods are Apr/2008-Mar/2018 and Apr/2003-Mar/2008, respectively), with an additional three years for warm-up. Nevertheless, we decided to include results only for the calibration period in this manuscript since the goal is to highlight that the underlying reason for poor results in simulated spatial patterns and annual cycles of variables other than Q are calibration-related decisions. We will include this justification at the beginning of section 4.3:

“We considered Apr/2005-Mar/2018 and Apr/1995-Mar/2005 as the calibration and evaluation periods, respectively (excepting the Futa River basin, where calibration and evaluation periods are Apr/2008-Mar/2018 and Apr/2003-Mar/2008), with an additional three years for warm-up. Nevertheless, we focus on the results for the calibration period since our primary goal is to diagnose pitfalls in calibration decisions regarding simulated spatial patterns and annual cycles of variables other than Q.”

Nevertheless, we will include a new Figure in the Supporting Information section (Figure R9) with Q-based metrics obtained for the evaluation period (2003-2008 for the Futa River basin, and 1995-2005 for all the rest). Overall, the results reveal for most cases little variations in Q metrics when modifying the regularization scheme, with two notable exceptions: KGE(1/Q) (third row) and the Cochiguaz River basin. Indeed, Cochiguaz is the basin that changes the most – compared to the calibration period – in terms of OF values, KGE(Q) and NSE(FDC(Q)), with more pronounced differences among regularization schemes when maximizing KGE(Q)+KGE(1/Q) and KGE(Q)+NSE(FDC(Q)). During the evaluation period, the choice of regularization scheme also affects considerably the results of KGE(Q)+KGE(1/Q) in the Colorado River basin, and regularizing K_{sat} when maximizing this OF yields a decay of Q metrics in the Cautín River basin.

We will also include some of the figures presented in previous responses that include evaluation results in the Supporting Information (section S.6).

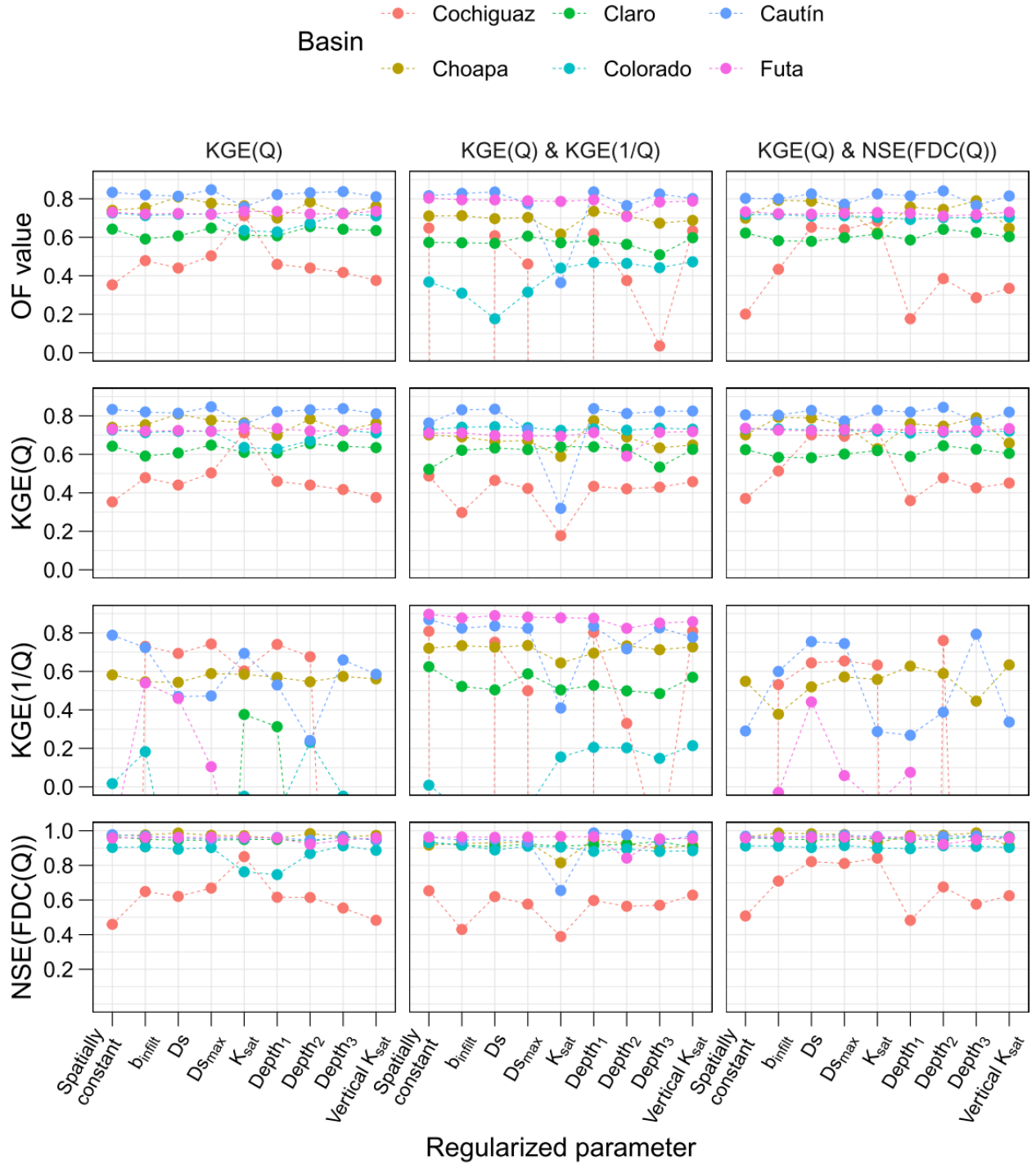


Figure R9: Streamflow performance metrics associated with the best parameter set obtained with spatially constant parameters (i.e., benchmark calibration, $\theta_a = 0$) and different regularization strategies during the evaluation period (April/2003-March/2008 for Futa, and April/1995-March/2005 for the rest). The columns show results for three different calibration objective functions. In each panel, the results for individual basins are displayed with different colors.

References

- Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., & Schaefli, B. (2020). Improving the Predictive Skill of a Distributed Hydrological Model by Calibration on Spatial Patterns With Multiple Satellite Data Sets. *Water Resources Research*, 56(1), 1–26. <https://doi.org/10.1029/2019WR026085>
- Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., & Stisen, S. (2018). Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrology and Earth System Sciences*, 22(2), 1299–1315. <https://doi.org/10.5194/hess-22-1299-2018>
- Khatami, S., Peel, M. C., Peterson, T. J., & Western, A. W. (2019). Equifinality and Flux Mapping: A New Approach to Model Evaluation and Process Representation Under Uncertainty. *Water Resources Research*, 55(11), 8922–8941. <https://doi.org/10.1029/2018WR023750>
- Koppa, A., Gebremichael, M., & Yeh, W. W. G. (2019). Multivariate calibration of large scale hydrologic models: The necessity and value of a Pareto optimal approach. *Advances in Water Resources*, 130(June), 129–146. <https://doi.org/10.1016/j.advwatres.2019.06.005>
- Kumar, R., Livneh, B., & Samaniego, L. (2013). Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme. *Water Resources Research*, 49(9), 5700–5714. <https://doi.org/10.1002/wrcr.20431>
- de Lavenne, A., Andréassian, V., Thirel, G., Ramos, M. -H., & Perrin, C. (2019). A Regularization Approach to Improve the Sequential Calibration of a Semidistributed Hydrological Model. *Water Resources Research*, 55(11), 8821–8839. <https://doi.org/10.1029/2018WR024266>
- López, P. L., Sutanudjaja, E. H., Schellekens, J., Sterk, G., & Bierkens, M. F. P. (2017). Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products. *Hydrology and Earth System Sciences*, 21(6), 3125–3144. <https://doi.org/10.5194/hess-21-3125-2017>
- Pokhrel, P., Gupta, H. V., & Wagener, T. (2008). A spatial regularization approach to parameter estimation for a distributed watershed model. *Water Resources Research*, 44(12). <https://doi.org/10.1029/2007WR006615>
- Pokhrel, P., & Gupta, H. V. (2010). On the use of spatial regularization strategies to improve calibration of distributed watershed models. *Water Resources Research*, 46(1), W01505. <https://doi.org/10.1029/2009WR008066>
- Pool, S., Fowler, K., & Peel, M. (2024). Benefit of Multivariate Model Calibration for Different Climatic Regions. *Water Resources Research*, 60(4). <https://doi.org/10.1029/2023WR036364>
- Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46(5), W05523. <https://doi.org/10.1029/2008WR007327>
- Shah, S., Duan, Z., Song, X., Li, R., Mao, H., Liu, J., et al. (2021). Evaluating the added value of multi-variable calibration of SWAT with remotely sensed evapotranspiration data for improving hydrological modeling. *Journal of Hydrology*, 603. <https://doi.org/10.1016/j.jhydrol.2021.127046>
- Tong, R., Parajka, J., Salentinig, A., Pfeil, I., Komma, J., Szeles, B., et al. (2021). The value of ASCAT soil moisture and MODIS snow cover data for calibrating a conceptual hydrologic model. *Hydrology and Earth System Sciences*, 25(3), 1389–1410. <https://doi.org/10.5194/hess-25-1389-2021>