Review reports on 7th of May

Editor message:

Dear Authors,
The revised version of your article is nearly ready for publication, but the reviewers have suggested a few additional changes. I agree with their recommendations and believe that making these changes will give the study greater visibility. Should you disagree with any of the suggested changes, please explain why clearly.

We would like to thank the editor and reviewers for their constructive feedback on Revision 1 of our manuscript, HESS_EGUSPHERE-2025-2478. Below, we present the comments from the two reviewers, followed by our responses to each comment (shown in blue font). We have incorporated their suggestions into Revision 2 through a revised title, text revisions, additional text and a revised figure. We believe these revisions have strengthened the presentation of our key findings in both the abstract and the main body of the manuscript.

Report #1

Submitted on 04th of Feb 2026

Anonymous referee #2

Overall, the manuscript has improved significantly since the initial version; however, it still requires clarification of the study objectives and scientific contributions in the abstract, possible changes to the title, and possibly an expansion or restructuring of the methodologies. How calibration is used to account for non-stationarity needs to be mentioned in the abstract – this is unclear without reading the methodology, but important to understand the paper's objectives and contributions. In addition, clarification of the methodological approach is needed, as inconsistencies in the time periods in the abstract and an unclear or missing description of which forcing data were used for the calibration and evaluation phases make it difficult to understand the connection to the future scenarios. This also relates to the manuscript's title, which is now better than in the first version, but it still seems to address two different objectives, which I don't think is the case.

We would like to thank the reviewer for his/her valuable comments. We have substantially revised the abstract, as presented and described below to better reflect the objectives, methods, and results presented in the manuscript.

Revised abstract
The use of conceptual hydrological models for projections of future freshwater resources is challenged by non-stationary climate conditions, as these conditions may affect whether models calibrated under historical climates are suitable for future scenarios. This study aims to (i) develop a framework for the parameterization of a conceptual hydrological model under non-stationary climate conditions and (ii) bias-correct, downscale and evaluate the performance of an 18-member ensemble of Regional Climate Models (RCMs) for simulating future streamflow. The framework was applied to generate streamflow projections for 38 mountain watersheds in the eastern Mediterranean island of Cyprus over the next decades (2030–2060) with the GR4J model. Six Nash-Sutcliffe Efficiency (NSE)- and Kling-Gupta Efficiency (KGE)-based functions and a composite scaled score were used for model calibration and validation across multiple 5-year calibration and 5-year validation periods (1980–2015). Climate non-stationarity was represented by differences in total precipitation between calibration and validation periods (ΔP=Pval-Pcal) using the differential split-sample test approach. The best-performing parameterization during drier periods (ΔP≤-5%) was obtained with an NSE objective function applied to square-root transformed streamflow, achieving NSE of 0.48, KGE of 0.54 and total streamflow bias of 9% during validation. This optimized model was selected for future streamflow simulations. Nine RCMs were excluded from the impact assessment because they underestimated the fraction of wet period precipitation (60-73% instead of 82%), resulting in streamflow biases up to 40% in the 1980-2010 reference period. The median of future projections for 2030–2060 shows a 6% reduction in precipitation and a 17% reduction in streamflow. In the worst case, reductions could reach 16% and 39%, respectively. Notably, during the driest years, streamflow reductions could reach 70% relative to historical dry years. Our findings suggest that terrestrial water resources in Cyprus may decrease significantly in the coming decades.

14: I agree with that statement. But it is not then clear why a calibration is pursued here, since calibration tailors a model to past observations. How does the presented methodology address this? In your title, you promise “calibration under non-stationary climate”: how is this achieved exactly? I assume you are using some kind of differential split sampling but it would be great if this could already be clarified in the abstract.

Thank you. We revised the abstract based on this comment and your following comments and the aspect of calibration under non-stationary climate is now made clear as presented in the following responses.

15: What model? It is rather confusing that you only start to use the specific name GR4J later in the abstract. And one could think that you are using multiple models here. I guess that this is not the case.

We added the name of the model used in Ln 18 of the revised manuscript.

"The framework was applied to generate streamflow projections for 38 mountain watersheds in the eastern Mediterranean island of Cyprus over the next decades (2030–2060) with the GR4J model".

20: This needs to be rephrased. You probably mean that a calibration with this specific objective function provided the best results. However, it is also unclear how "best performance" is quantified. Also, "drying trends" need to be quantified. Please provide concrete numbers here.

Thank you. We rephrased the sentence according to your recommendation in Ln 22-25.

"The best-performing parameterization during drier periods (ΔP≤-5%) was obtained with an NSE objective function applied to square-root transformed streamflow, achieving NSE of 0.48, KGE of 0.54 and total streamflow bias of 9% during validation".

21: What does "this method" exactly refer to? Are you really presenting something new? If that is the case, the method deserves a name to clarify it. But also a short statement on what is new compared to the existing methods. So far, it sounds rather traditional to calibrate a model with multiple objective functions and then test which provides the best results. The transferability is thus unsurprising, since modeling teams have been doing this for quite a while. I guess you are essentially applying a modified DSST here, right.

We changed the "comparative method" and "this method" and added some more descriptive sentences for our methods in Ln 20-25.

"Six Nash-Sutcliffe Efficiency (NSE)- and Kling-Gupta Efficiency (KGE)-based functions and a composite scaled score were used for model calibration and validation across multiple 5-year calibration and 5-year validation periods (1980–2015). Climate non-stationarity was represented by differences in total precipitation between calibration and validation periods (ΔP=Pval-Pcal) using the differential split-sample test approach."

We agree that the general concept of comparing multiple objective functions and model calibration using DSST is not entirely new. Our contribution lies in a systematic evaluation of simulated streamflow, both for model calibration with different objective functions and for non-stationary climate conditions between calibration and validation, as mentioned in more detail in Ln 101-105. Our model experiments consist of an "extended" DSST with 14 model parameterizations and 182 model validations for different changes in precipitation, as explained in Section 2.1.2.

23: Now, this is the part where it gets confusing. It is great that you are using an ensemble of RCMs, but why do you simulate a different period (till 2010 instead of till 2015 – see line 19)? And in addition, you evaluate the ensemble, which is very cool, but it is unclear to me how streamflow bias was estimated.

Thank you. We use the longer period 1980-2015 for calibration and evaluation for the reason of including as many hydrological years as possible for examining the calibrated model transferability under different climate conditions. Streamflow observations for all studied watersheds are available from the monitoring authorities until 2015 (Ln 227-230). For the climate change impacts, we used the shorter reference period, 1980-2010, for the reason of matching the reference period length to the climate projections period length (2030-2060) (Ln 251-253).

The method followed to compute the streamflow bias is explained in section 2.4 "Selection of RCMs for the forcing of future streamflow simulations" and in Ln 269-276.

The revised abstract explains now for which purposes the periods 1980-2015 and 1980-2010 are used.

26: What does accurately mean?

We revised the sentence in Ln 25 in order to shorten the abstract, so that the sentence with "accurately" is also removed.

"Nine RCMs were excluded from the impact assessment because they underestimated the fraction of wet period precipitation (60-73% instead of 82%), resulting in streamflow biases up to 40% in the 1980-2010 reference period"

26 and following: So you are using the calibrated hydrological model here? If that is the case, say that!

Thank you. Yes, we use the calibrated model for future projections. We modified Ln 25-32 to make this clear.

"The best-performing parameterization during drier periods (ΔP≤-5%) was obtained with an NSE objective function applied to square-root transformed streamflow, achieving NSE of 0.48, KGE of 0.54 and total streamflow bias of 9% during validation. This optimized model was selected for future streamflow simulations."

I suggest modifying the abstract to 1) first talk about the issue which needs to be defined better, 2) to then state that you use an ensemble of regional climate models that you evaluate, 3) and that you then calibrate a hydrological model with multiple objective functions with (streamflow as a target? This is also unclear) and select the best performing one for an ensemble simulation of the future. I think if this is what you are doing, you could also change your title to "Selective ensemble of regional climate models forcing a calibrated hydrological model reveals future decreasing streamflow conditions in Cyprus," for example. It doesn't have to be this title, but I think it would be helpful to make it clearer that this is a connected effort to provide accurate simulations. The revised title still sounds like you pursued different objectives. This also remains unclear in the methodology of the paper. What forcing is used in the calibration and evaluation phase, and how is it connected to the simulations of future scenarios?

Regarding your points,
1) We modified the abstract substantially and we believe the objectives and methods discussed in the paper are now reflected more clearly in the abstract.

2) We modified Ln 15-16 to make clear that an RCM ensemble is evaluated.

" **...** and (ii) to bias-correct, downscale and evaluate the performance of an 18-member ensemble of Regional Climate Models (RCMs) for simulating future streamflow ."

3) We modified Ln 16-18 to make clear that the calibration target variable is streamflow.

"The framework was applied to generate streamflow projections for 38 mountain watersheds in the eastern Mediterranean island of Cyprus over the next decades (2030–2060) with the GR4J model."

We also added that the best calibrated GR4J setup was used for the ensemble of future projections in Ln 24-25 as explained in our response in the previous comment.

Title: We appreciate your title suggestion for making our manuscript title more coherent.
Our revised title is :
"Assessing future streamflow in Cyprus through hydrological model calibration under non-stationary climate and Regional Climate Model ensemble selection"

Regarding the forcing used in the calibration and evaluation phase, and how is it connected to the simulations of future scenarios, we present the observations-based forcing for the calibration and evaluation of GR4J in Sections "2.2 Study area and observational data" and "2.1.2 Selection of objective function under changing climate conditions", and the RCM-based forcing in Section "2.4 Selection of RCMs for the forcing of future streamflow simulations". We believe that the details added in the revised abstract match now the forcing datasets explained in these sections.

29: How do you define historical dry years?
We define historical dry years as the "two consecutive driest and wettest years as well as the five driest and five wettest years overall" in Ln 281-283.

Fig. 2: Maybe I missed it in the text, but can you speculate why KGElog shows a higher CSS with higher P when delta-P is >15%, but not when delta-P<-5%, and why NSEsqrt instead shows this relationship?

Thank you for pointing out this interesting finding in the figure that deserves more attention. For this reason, we added few sentences to give possible explanations to the patterns observed in Figure 2 in Ln 318-328.

"The patterns in Figure 2 may be interpreted in relation to both the streamflow transformation and the formulation of the objective functions. The relatively stable or improving performance of the model calibrated with NSElog and KGElog with increasing watershed wetness suggests that logarithmic transformation reduces the influence of the larger and more frequent peak flows, characteristic of wetter catchments. This tendency is most apparent under wetter validation conditions (ΔP > 15%). However, this advantage is not observed during drying validation periods (ΔP ≤ −5%), possibly because reduced-flow conditions increase the relative importance of bias and variability errors between simulations and observations. Since KGE explicitly incorporates correlation, bias, and variability components, KGE-based calibration may be particularly sensitive to these changes under drying conditions, even when logarithmic transformation is applied. In

contrast, NSEsqrt appears more robust across precipitation regimes. The square-root transformation has a weaker effect in attenuation of peak flows, compared to the logarithmic, and thus it may retain a more balanced representation between moderate and high flows during evaluation across both wetting and drying conditions. In contrast, calibrations with with KGEsqrt show consistently weaker performance across precipitation conditions. These findings suggest that the interaction between transformation type and objective-function structure differs between NSE and KGE formulations, with NSEsqrt providing a more stable compromise across contrasting hydroclimatic conditions. "

Report #2

Submitted on 29 Apr 2026

Anonymous referee #1

I have been through the authors revisions, and I am by and large satisfied with their revised manuscript subject to comments and suggested amendments below.

We would like to thank the reviewer for his/her valuable comments, which are addressed below.

Introduction
Ln 105 Revise below sentence.
"The specific objectives are: (i) to develop a method for evaluating the performance of a hydrological model when calibrated with different objective functions under a changing climate."
It seems to suggest that it was calibrated under a changing climate? Are you not looking to evaluate a method for evaluating the robustness of a hydrological model for undertaking future climate simulations? Currently objective seems a bit muddled and hard to follow.

Thank you for your comment. We revised specific objectives (i) and (ii) following your suggestion in Ln 115-122.

"The specific objectives are: (i) to develop a framework for the parameterization of conceptual hydrologic models for robust streamflow simulations under non-stationary climate conditions, (ii) to bias-correct, downscale and evaluate the performance of an 18-member RCMs ensemble from CMIP5 models for streamflow simulations, and (iii) to apply the optimized hydrological model parameterization with selected RCMs for 38 mountain watersheds in Cyprus, and assess the mid-term future (2030-2060) impact on the island's water resources."

Data and methods
Ln 122 "Percent bias (PBIAS) was also included to provide an independent measure of total volume error (Moriasi et al. 2007; Coron et al. 2012)."
This is ambiguous. Was PBIAS included in the calibration ie as a penalty to the objective function score if PBIAS was outside an allowable range? Or was it just included as an evaluation metric. If the latter, then this is already said in Line 125.

We removed the sentence in Ln 122 and revised the sentence in Ln 123.

"Percent bias (PBIAS) was used as a seventh evaluation measure to account for total volume error (Moriasi et al. 2007; Coron et al. 2012)."

Ln 170 A comparison of 5 year and 18 year calibration scores is meaningless in terms of evaluating the robustness of the parameter sets calibrated on 5 years of data. I would expect calibration scores for 5 years of data to the be same if not higher than calibration over 18 years of data. The calibration score is simply a measure by which the model can be made to fit the data. I could only have 1 year of data and get an equally high calibration score, but my model is likely to be far less robust. To actually test the robustness you have to validate against an independent dataset. My original concern remains given the literature suggests about 10 years of data is required?

Thank you. Our model calibration experiments were evaluated based on the results of multiple independent 5-year validation runs as explained in Ln 160-162. We added the model evaluation results of the 5-year against 18-year calibration during an independent validation period in Ln 167-173.

"Although longer calibration periods are often recommended, this study adopted a 5-year calibration and validation window as a compromise between capturing interannual climate and hydrological variability typical of Mediterranean conditions and maximizing the number of sub-periods for comparison. A comparison of the 5-year against an 18-year calibration was made showing very similar median performance across 38 watersheds during calibration (NSE: 0.82 vs. 0.83; KGE: 0.89 vs. 0.88) and during an independent 17-year validation (NSE: 0.62 vs. 0.65; KGE: 0.57 vs. 0.59). During validation,32 out of 38 watersheds exhibited NSE differences <0.1 and 28 out of 38 exhibited KGE differences <0.1."

Ln 249 Here and elsewhere. My understanding is that RCP8.5 is no longer the business-as-usual pathway? Authors should revisit their framing for why they used RCP8.5 and discuss implications of this more contemporary thinking around which projections are most likely in terms of their results. See Hausfather Z and Peters GP (2020) Emissions–the 'business as usual' story is misleading. Nature 577(7792), 618-620.

Thank you for highlighting the evolving perspective regarding the interpretation of RCP8.5. We have revised our characterization of RCP8.5 by removing the "business-as-usual" framing and clarifying why it was selected as the focus of our future projections (Ln 256-259).

"The RCP8.5 scenario, corresponding to the upper range of the projected global-mean surface temperature increases (Meinshausen et al., 2011) was selected to quantify the effects of the least favorable among the plausible climate change scenarios (Pedersen et al., 2021) on surface water resources."

Figure 2 legend (which shows markers) doesn't correspond to the lines in plots?
We corrected the legend to include dashed lines that correspond to the linear trend between the score (CSS) and precipitation.

Ln 350 "This finding suggests that understanding how different objective functions lead to different model outputs, in relation to the input model parameters and interannual variations in precipitation forcing, is important. "
The authors present these result but the read is then just kind of left hanging, with no 'so what'. Essentially the authors have done the analysis suggested by the second reviewer but have not provided any interpretation of the results of discussion of what the implications might be for their study. I understand the paper is already long and this is potentially a substantial piece of work but it probably needs to be rounded off a bit better, at the moment it reads like a very obvious and unsatisfying loose end.

For example, the X2 value is quite negative, very negative in some catchments. I understand this to essentially be a by GR4J since it tends to underestimate ET in dry conditions via the production store.
Hughes, Potter, Zhang (2015) Is inter-basin groundwater exchange required in rainfall–runoff models: The Australian context. 21st International Congress on Modelling and Simulation

We have modified the paragraph and added a possible interpretation of our results linked to the land use conditions of the studied watersheds (Ln 368-381).

"The optimized model parameters derived using different objective functions and calibration periods were plotted against the average precipitation of the respective calibration periods (Figure A1). Although the model parameters do not directly correspond to measurable watershed properties, their calibrated values can provide insight into integrated watershed behavior. In particular, parameter X2 (groundwater exchange) calibrated values range from negative to near zero. In GR4J, negative X2 values imply a loss from the water balance of a watershed. This behaviour may reflect processes not explicitly represented by the model, such as groundwater abstractions for irrigation. The three watersheds with the most negative X2 values have a larger fraction of agricultural land cover (43%-48%), whereas the five watersheds with near-zero X2 have a larger fraction of forest cover (74%-99%) (Sofokleous et al. 2023). This pattern suggests that calibration compensates for water losses not explicitly represented in GR4J. In addition, the results in Figure A1 also show that, for a given watershed, parameter values generally vary more between different objective functions than between different calibration periods. Therefore, understanding how different objective functions lead to different model outputs is important for hydrological model applications in areas with diverse watersheds conditions and high climate variability."

I couldn't find anything in the study area description to help me understand how valid this inter-basin groundwater transfer would be in the catchments examined in this study (ie are they likely to be open or closed catchments based on their geology). See my comment further down about the applicability of Gr4J in modelling runoff over extended dry periods.

Thank you. We added some information in Section "2.2 Study area and observational data" about the geology of the study area that could help draw conclusions on the calibrated model parameters (Ln 198-204).

“Geologically, Troodos is constituted by an ophiolite complex with faulted and highly fractured rocks, especially the gabbro on the upper hillslope, forming fractured aquifers or aquifer systems favoring infiltration (Udluft et al., 2006). Soils in Troodos have a stony gravelly texture and a high variability in soil depth from very shallow (0–10 cm) up to about 100 cm (Camera et al., 2017). Christofi et al. (2020) used isotope and hydrogeochemical sampling and modelling to identify regional groundwater flow through the diabase and basal-group units of the Troodos Fractured Aquifer, providing evidence of groundwater flow across surface watershed boundaries in the Troodos Mountains.”

Ln 411-412 Would be more logical to speak to the rainfall results first then streamflow.

Thank you. We changed the order of the presentation of our results in the first paragraph of Section “3.4 Future projections of water resources”.

“Precipitation in 2030-2060 is projected to decrease by 16% according to the driest model and by 6% according to the median model, relative to the 562 mm·y-1 over the 38 watersheds for the reference period (Table 3). Total streamflow of the 38 watersheds is projected to remain the same in the best model case or decrease up to 39%, in the worst case, relative to the 1980-2010 reference period value of 173 $Mm^3 \cdot y^{-1}$.”

Ln 576-580 “The emphasis was placed on multiple RCMs rather than multiple hydrological models, as previous research has shown that variability in rainfall–runoff model outputs is greater for mid- to high-flow and mean annual flow conditions when runoff projections are based on a single rainfall–runoff model combined with multiple climate models, rather than the reverse (Teng et al. 2012; Petheram et al. 2012).?”

"Rather than the reverse" does not make sense to me? The Teng and Petheram papers demonstrated that yes different GCM result in more variability in runoff than different RRM for mid-to-high flows. But they also showed that for low flows variability in runoff due to different RRM is higher than due to different GCM. This is not what “rather than the reverse” means to me. Note don’t feel you need to include these references I was just pointing out that there are studies (looking at low flows) that dispute the authors statement in the original paper, which implied this was the case for all flows.

Thank you. We removed the “rather than the reverse” and revised the sentence to avoid overgeneralizing across flow regimes (Ln 604-608).

“The emphasis was placed on multiple RCMs rather than multiple hydrological models, as previous research has shown that for mid- to high-flow and mean annual flow conditions, variability in rainfall–runoff model outputs is greater when runoff projections are based on a single rainfall–runoff model combined with multiple climate models (Teng et al. 2012; Petheram et al. 2012).”

Ln 584-586 “A single hydrological model, GR4J, was used, as the particular model is a well-established conceptual model structure for streamflow simulations across a wide range of hydroclimatic conditions.”
Yes it has been widely used, however, some authors are note deficiencies in GR4J in its ability to accurate simulate runoff over extended dry periods. This would seem to be particularly relevant to this paper in modelling projected dry future climates? Clearly the authors are not going to redo their modelling using GR7J or other RMM, as there are still important learnings from this paper. However, the authors should read the papers for which I provide links below (and other similar papers), and they should appropriately caveat this statement. In fact the authors should probably caveat the paper more broadly based on the below and similar papers, and briefly discuss what implications these may have for the results of this paper.

Grigg and Hughes (2018) Nonstationarity driven by multidecadal change in catchment groundwater storage: A test of modifications to a common rainfall–run-off model. Hydrological Processes, 32, 24.

Hughes, J., Potter, N., Zhang, L., & Bridgart, R. (2021). Conceptual Model Modification and the Millennium Drought of Southeastern Australia. Water, 13(5), 669. https://doi.org/10.3390/w13050669

Keirnan Fowler, Wouter Knoben, Murray Peel, Tim Peterson, Dongryeol Ryu, Margarita Saft, Ki-Weon Seo, Andrew Western. (2020) Many Commonly Used Rainfall-Runoff Models Lack Long, Slow Dynamics: Implications for Runoff Projections. WRR, 56, 5.

Indeed, this limitation of rainfall-runoff models should be mentioned for climate change applications involving extended drying climate patterns. We added some discussion sentences to discuss the important issue raised in Ln 616-621.

“ However, uncertainties also arise from the hydrological model structure used in impact studies. A particular limitation of “bucket-type” rainfall–runoff models, such as GR4J, is their tendency to underestimate multi-year drought conditions due to finite storage, which limits their ability to represent sustained groundwater decline (Fowler et al., 2020). The use of model parameters calibrated during observed dry conditions with trends similar to the projected climate, as suggested in

this study, and structural modification of these models to account for catchment memory (e.g. Grigg and Hughes, 2018; Hughes et al., 2021) could help reduce this underestimation."