Reply to Referee 1 in black :

Authors: We thank the Referee for the detailed assessment of our study. Following the comments and recommendations, we have revised the way the models are presented: we now present Model 1 without a snow routine, followed by Model 2 with a snow routine. This validated Model 2 with a snow routine is then used to test two scenarios involving an increase in air temperature. The two scenarios enable us to conduct sensitivity tests on the response of spring discharge to temperature variations in a karst aquifer with seasonal snow. In the revised version of the article, we have modified the results presented for the temperature increase scenarios: we now show 5,000 simulation results, and can thus provide an uncertainty assessment of the results. We therefore believe we have improved the study and addressed the main weakness identified by Referee 1. The article now appears to us to be more robust from a methodological standpoint, which in turn lends robustness to the reported results on the impact of air temperature variations on the discharge of the karst springs studied.
We would also like to take this opportunity to add Dr. Vianney Sivelle as a co-author (CNRS - France). His contribution was essential in addressing the points raised by the referee.

Below, we provide detailed responses to the Referee's comments. In this reply, we have numbered each comment using the notation R1 (Referee 1) followed by the comment number.

R1.1: As already mentioned, the current study design seems to be a bit shallow. While I agree that the study of temperature changes and their impacts on catchment dynamics is relevant, this seems to be done a bit superficial right now. The use of a very conceptual model does not really allow to expand the study to analyse model system states or related aspects. However, what I would think could be interesting in this context is an uncertainty assessment of the model (results). As described in line 286, 10,000 parameter sets yield simulation results that satisfy the objective function criteria. These could be used to conduct the analysis regarding the effect of temperature changes. This would result in a range of potential model reaction behaviour under changing temperatures, letting us make more robust interpretations and deductions from the results. I think this would greatly enhance the study. In the same way, different model structures and their impact on simulated discharge under temperature shift could be included in the analysis — as you mentioned that there are different structures that show equally good model performance.

Authors: As noted by the reviewer, the model optimizes all parameters using a quasi-Monte Carlo approach (n parameter sets), which gives access to n results. In the initial version of the paper, we used the parameter set giving the highest objective function value as the calibrated model ($W_{obj\_max}$). Following this suggestion, we have revised the paper to incorporate the full ensemble of n simulation results for the temperature change scenarios, providing an uncertainty assessment of the model results and giving more robustness to the predictive discharge, including the range of results between 5-95th percentile.
Using the KarstMod software, we simulated n parameter sets for calibration and validation periods, benefiting from a new beta version of the software that allows the use of n sets of calibrated parameters to run simulations with modified input conditions (e.g., a change in temperature). We retain the 5,000 best parameter sets, for which the objective function exceeds 0.85. Results are presented in time series graphs and violin plots (Matplotlib-Python), and figures have been revised accordingly, also incorporating the visualization improvements suggested by Referee 2.
To illustrate the consistency between both approaches, we focus on August, the critical low-flow month when water demand is highest. In the original manuscript, monthly mean August discharge was reported as 1.19 $m^3 \cdot s^{-1}$ for Model 2, decreasing to 0.86 $m^3 \cdot s^{-1}$ under T+2°C (-28%) and 0.67 $m^3 \cdot s^{-1}$ under

T+4°C (-44%). In the revised analysis, the ensemble median values for August are 1.13 ± 0.29 $m^3 \cdot s^{-1}$ for Model 2, 0.78 ± 0.25 $m^3 \cdot s^{-1}$ under T+2°C (−31%) and 0.59 ± 0.22 $m^3 \cdot s^{-1}$ under T+4°C (-48%), where uncertainty is expressed as the Median Absolute Deviation (MAD) across the 5,000-member parameter ensemble. The 5-95th percentile ranges are 0.46-1.77, 0.24-1.36 and 0.15-1.15 $m^3 \cdot s^{-1}$ respectively.
The relative decreases between scenarios are consistent between the two approaches (-28% and -44% vs. -31% and -48%). The projected reduction in August discharge under both warming scenarios is therefore a robust signal across the full parameter space.
Regarding the suggestion to compare different model structures, we agree this is an interesting perspective. However, testing and comparing numerous model structures requires extensive analysis that goes beyond the scope of the present study. The chosen model structure (based on preliminary recession curve analysis) is a standard karst model that adequately represents the fast and slow flow components typical of the Dévoluy system. The primary objective of this study is to accurately reproduce the observed discharge at this specific catchment, and to assess the effect of seasonal snow under warming temperature scenarios. The revised manuscript now includes 5000 simulation results for the temperature increase scenarios, providing a robust uncertainty assessment of the projected hydrological changes. The influence of model structure on the results could be addressed in future work.

R1.2: Also, I dont really see the sense behind the split in model 1 and model 2 in the current study design. If model 1 is without snow routine, meaning that all precipitation occurs as rainfall, then why is a model 2 necessary with snow routine deactivated? From my understanding, the study only needs one model without snow routine and one model with snow routine (both including potential parameter sets / model structures that show satisfactory model performance) to assess the impact of snow inclusion on model performance. This would enhance clarity regarding the methodology.

Authors: We agree the recommendation of the reviewer. We believe that the paper will be clearer by presenting only one model without snow routine and one model with snow routine, and then 2 scenarios of air temperature shift. We will remove Model_2_without snow routine. The two scenarios (Model 2 T+2°C and Model 2 T+4°C) are enough to show the case with a large decreasing amount of snowfall.

R1.3: Lines 74 to 77: I think this is more methodology rather than being relevant for the introduction.

Authors: These sentences focus on a brief presentation of the case study. We think it is few basic information needed by the readers to catch the interest of this Alpine case study, to explore the topic of snow and karst discharge.

R1.4: Line 180: I dont think this figure is necessary in the current form. Subfigures a and c really dont offer anything that needs to be depicted as a figures. Subfigures b and d are debatable, if they are really necessary, as the snow depth is shown later within comparisons as well as the discharge. The only unique information here is the discharge for the Souloise river, which is not used later if I am correct?

Authors: We agree that this figure could be simplified by deleting subfigure (c). We chose to keep (a) precipitation and temperature to illustrate the range of daily values in this case study in France, (b) discharge at Gillardes springs and Souloise river because comparison between river and spring is used to show the predominance of the karst discharge, (c) snow observation demonstrates that this mid altitude karstic catchment is subject to seasonal snow and seems important for us to be shown. We can also put this figure in supplementary material if the reviewer and editor want to shorten the length of the paper.

R1.5: Lines 204 to 208: Are [L], [L/T], or [/T] supposed to be the units for the corresponding values? Are those something like $m^3$, litres, $km^2$?

Authors: This corresponds to the physical dimensions of the various variables to allow kind of generalization of the model regarding the input data provided by the user. To avoid potential confusion, we will replace L by Length, T by Time.

R1.6: Line 222: I would suggest writing things like Wobj or anything like that as some sort of variable: $W_{obj}$, $Wobj_{min}$, $E_{min}$, $P_{sr}$.

Authors: We will replace Wobj as $W_{obj}$, Emin as $E_{min}$ and other variable as you suggested.

R1.7: Line 234: I think to say that you only using one single evaluation crtieria is similar to a multi-objective criteria calibration is difficult, because it's not true. You still only rely on the discharge, so no other aspects like snow height or ETa are explicitly evaluated, which would be a real multi-criteria evaluation.

Authors: We agree that referring to a single composite performance criterion as being comparable to a true multi-objective calibration may be misleading. Although KGEnp integrates several statistical components (correlation, bias, and variability), the calibration in our study is indeed based solely on discharge observations and does not include additional state variables such as snow storage or actual evapotranspiration.
We propose to revise the manuscript accordingly to clarify that KGEnp is a composite metric applied to discharge only, and we will change wording suggesting equivalence with a true multi-criteria or multi-variable calibration.

R1.8: Lines 237 to 242: These "soft-criteria" mentioned here are surely important, but the calibration (and evaluation) of model parameterisations is only done on the discharge by the KGEnp. Why are those aspects mentioned here not incorporated into the calibration/evaluation?

Authors: The spring discharge is correctly modeled (using the KGEnp objective function) by automatically adjusting the parameters that influence flow dynamics (k coefficients) and also the parameters for snow accumulation and melt (snow routine parameters). The soft criteria we list are proposed by the KarstMod software and we mention them as part of a "good practice" approach to modeling.

R1.9: Lines 247 to 265: Why is the snow routine introduced after model optimization? I think it would make more sense if the order is following a more logic way. First the study area, then the data, then the model and how it works, then how its is calibrated and then how the further study is conducted.

Authors: We agree with this suggestion and we will switch part 3.2 and 3.3 to present first the snow routine and then the model optimization.

R1.10: Lines 269: I feel like the figure would benefit from a legend that shortly points out what each abbreviation stands for. Currently it is in the caption, but that makes it less easy to grasp.

Authors: We choose a similar representation to the one regularly used in the literature dealing with KarstMod rainfall discharge model and its applications (Çallı et al., 2022; Mazzilli et al., 2019; Sivelle et al., 2019) . To keep the figure clear, we suggest keeping the abbreviations explained in the caption.

R1.11: Lines 271 to 275: I think this paragraph can be deleted. Lines 276 to 277: This is more suited for the introduction. No need here.

Authors: We agree with the reviewer's comment. Lines 271–275 have been removed. Regarding lines 276–277, we agree that this sentence is better suited for the introduction and have moved it accordingly. We will simplify by removing the lines mentioned and start the section directly with the modeling strategy.

R1.12: Lines 286 to 288: Why is only the "best" model retained if there are several parameterizations that seem to be fulfilling the criterion — and therefore show sufficiently good model behaviour?

Authors: We believe this comment is a key comment in this review. As answered in the introductory comment R1.1, we propose revising the paper by adding the results of the n=5000 sets of calibrated parameters applied to the 2 warming scenarios. We have run the 5000 simulations with modified input conditions (i.e., temperature shift T+2°C or T+4°C). All the results satisfy Wobj (KGEnp) > 0.85. Graphical results will show the range of probable discharge simulation, highlighted by specific curves (e.g. median, percentile 5 and 95). The previously “best model” (for Wobj max) will be only one of the possible result. We will so be able to assess the uncertainty of the modeling results, and give more robustness to the predictive discharge.

R1.13: Also, what does "Cross calibration validation tests did not improve performance" mean here? From my understanding and experience, the sense of a cross-calibration would here be to assess, if model performance is highly influenced (dominated) by a single year, for example.

Authors: Concerning the cross-calibration tests, we agree that our initial sentence was unclear. We performed a split-sample test by exchanging the calibration and validation periods to assess whether model performance was strongly dependent on a specific time interval (validation from September 2015 to September 2016 and calibration from September 2016 to September 2019). The resulting Wobj values remained comparable between calibration and validation configurations (higher than 0.8), indicating that model performance is not dominated by a particular year or sub-period and supporting the temporal robustness of the calibrated parameter sets. We will modify the text.

R1.14: Lines 320 to 324: kinda a repetition of what is mentioned in lines 310 to 313.

Authors: We agree that this section contained some redundancy. In the revised manuscript, we will condense the two paragraphs to avoid repetition and to present the information only once in a clearer and more concise manner.

R1.15: Lines 324 to 325: Expalantion of positive and negative shift makes no sense in this formulation. It should be formulated more as a relative value that is negative for heights with greater values than the reference and positive when heights are lower than the reference.

Authors: We will change the formulation of the sentence according to your recommendation.

R1.16: Lines 332 to 338: This can be combined with the explanations before to shorten on this, as it seems trivial.

Authors: We will shorten the explanation on the Ts parameter, highlighting its role and how it should be parametrized.

R1.17: Lines 339 to 342: You said those alternative structures did not improve model performance. Does that mean, those are equally feasible? Why are those then not incorporated into evaluation? Model structure uncertainty could be well assessed here and would help increase the robustness of the model result’s interpretation.

Authors: The current structure of the model is based on the identification of the main recession coefficients in the spring discharge time series before the rainfall-snow-discharge modeling. Once the main flow components (very fast flow, fast flow and slow flow) have been identified and treated as compartments, the purpose of the study was not to test several structures. The sentence line 339 was unclear. We should keep in mind that others structures exist, and can be tested, but it was out of the scope of the study. We will modify the text of the paper to address these possibilities in future work.

R1.18: Line 356: Isn't the comparison of the model with and without snow module the relevant method to assess the impact of snow incorporation on model performance, rather than the inclusion of different temperature changes?

Authors: We refer to our previous response regarding the removal of Model 2_SRoff on comment #R1.2 As explained, this configuration corresponds to a boundary case in which temperature changes would prevent snowfall entirely. To simplify the manuscript and clarify the presentation of the simulation workflow, we have removed this model from the study.

R1.19: Lines 369 to 373: This is more suited for model description.

Authors: We have deleted the model 2_SRoff accordingly to your recommendation so we deleted theses sentences.

R1.20: Lines 374 to 377: Delete.

Authors : We will remove this paragraph and cite in the introduction the literature about recent work about Explore 2 data using lumped models.

R1.21: Line 380: The figure should be updated accordingly with the changes made to the methodology.

Authors: As we decided to remove the use of the model_2SRoff we will modify the figure accordingly.

R1.22: Line 389: The table can probably be moved to the supplementary.

Authors : We chose to keep this table in the main text, as it provides essential information about the model parameters, their calibration ranges, and optimized values. This information is key for the reader to directly assess the calibration results and to evaluate the physical consistency of the optimized parameter sets. Moving it to the supplementary material would reduce the readability of the results section. However, following the suggestion of Referee 2, we will add additional efficiency metrics to complement the table.

R1.23: Line 416: Figure numbers are wrong through the whole manuscript from here. Please check.

Authors: There was a problem with figure numeration from there, we will fix it.

R1.24: Line 425: Please check sentence syntax.

Authors: We will modify the sentence to make it clearer.

R1.25: Line 428: The model can not provide a more robust performance criteria. Semantically not logic.

Authors: The sentence was not clear. We compare results and performance criteria from Model 2 to Model 1 (initial model without snow routine). This is why we said that the value of the performance criteria is more robust, higher. We modified the sentence to make is clear and logic.

R1.26: Line 443: What sense has a model here that is not validated?

Authors: In section 4.1.1 we demonstrate that Model 1 is not valid for the Dévoluy catchment, as it fails to correctly reproduce discharge during periods influenced by seasonal snow. Presenting temperature change projections with an invalid model would therefore be meaningless. Additionally, as discussed in our response to R1.2, we will remove Model_2SRoff and section 4.2.1, as the two warming scenarios (Model 2 T+2°C and Model 2 T+4°C) are sufficient to illustrate the expected decrease in snow-driven discharge under future climate conditions.

R1.27: Lines 512 to 513: If these temperature variant simulations are not to be understood as predictive forecasts, why is it framed like that through the manuscript?

Authors: We thank the reviewer for raising this point. We acknowledge that the framing of the temperature-variant simulations may have been ambiguous throughout the manuscript. These simulations are not predictive forecasts based on projected climate data, but sensitivity tests assessing the impact of temperature increases on karst aquifer dynamics using observed inputs modified by a uniform temperature offset. We will revise the abstract and introduction to explicitly frame these scenarios as sensitivity analyses, as also requested by Referee 3, and will clarify that this represents a first-step approach that could be further improved by incorporating fully predicted input datasets.

R1.28: Lines 590 to 592: Does the fact, that the scaling parameter for the catchment had to be expanded to 191 km² instead of the actual 150 km² not imply that underground exchanges or a differing aquifer catchment size are very likely possible?

Authors: We do not consider that a larger catchment size or significant exchanges with other aquifers are plausible. That is why we consider that this overestimation of the catchment size is due to a bias in precipitation in input data. However, this point was not sufficiently explained in the initial version of the manuscript, particularly in the Case Study presentation section. The Dévoluy system is a perched karst aquifer that is clearly disconnected from other potential aquifers due to its topographic position and geological setting, as shown in Fig. 1. The size of the catchment is therefore a well-defined parameter. We will modify the manuscript to ensure that this aspect is now clearly explained.

Minor comments:

R1.29: Line 126: main instead of "mains"

R1.30: Line 171: are rare and make

R1.31: Line 193: hydrological

R1.32: Line 207: the reference unit length is fixed Lines 249 to 251: Units in []-brackets, multiply dots wrongly formatted.

R1.33: Line 268: Input liquid precipitation stands for rainfall?

Authors: Yes.

R1.34: Lines 308 to 310: Formal language.

R1.35: Line 330: Range

R1.36: Line 331: unit formatting

R1.37: Line 359: "climate models already show warming"

R1.38: Line 364: ambiguous formulation regarding the temperature definition. Do you mean you use the temperature recommendations to define the simulated temperature increases?

Authors: As also suggested by Referee 3, we clarify that the French “TRACC” dataset consists of a set of regional climate trajectories. From these trajectories, we extract the temperature perturbations applied in our simulations. In other words, the temperature shifts imposed in the model follow the guidance provided by TRACC for plausible warming scenarios. The manuscript will be revised to make this more explicit and remove any ambiguity in the original formulation.

## R1.39: Line 486: comma missing

Modified figures :

Please find below the modified figures and the caption. Here the 5000 simulations satisfying $W_{obj\_min}$ = 0.85 in Model2 were kept and used for the temperature scenarios.

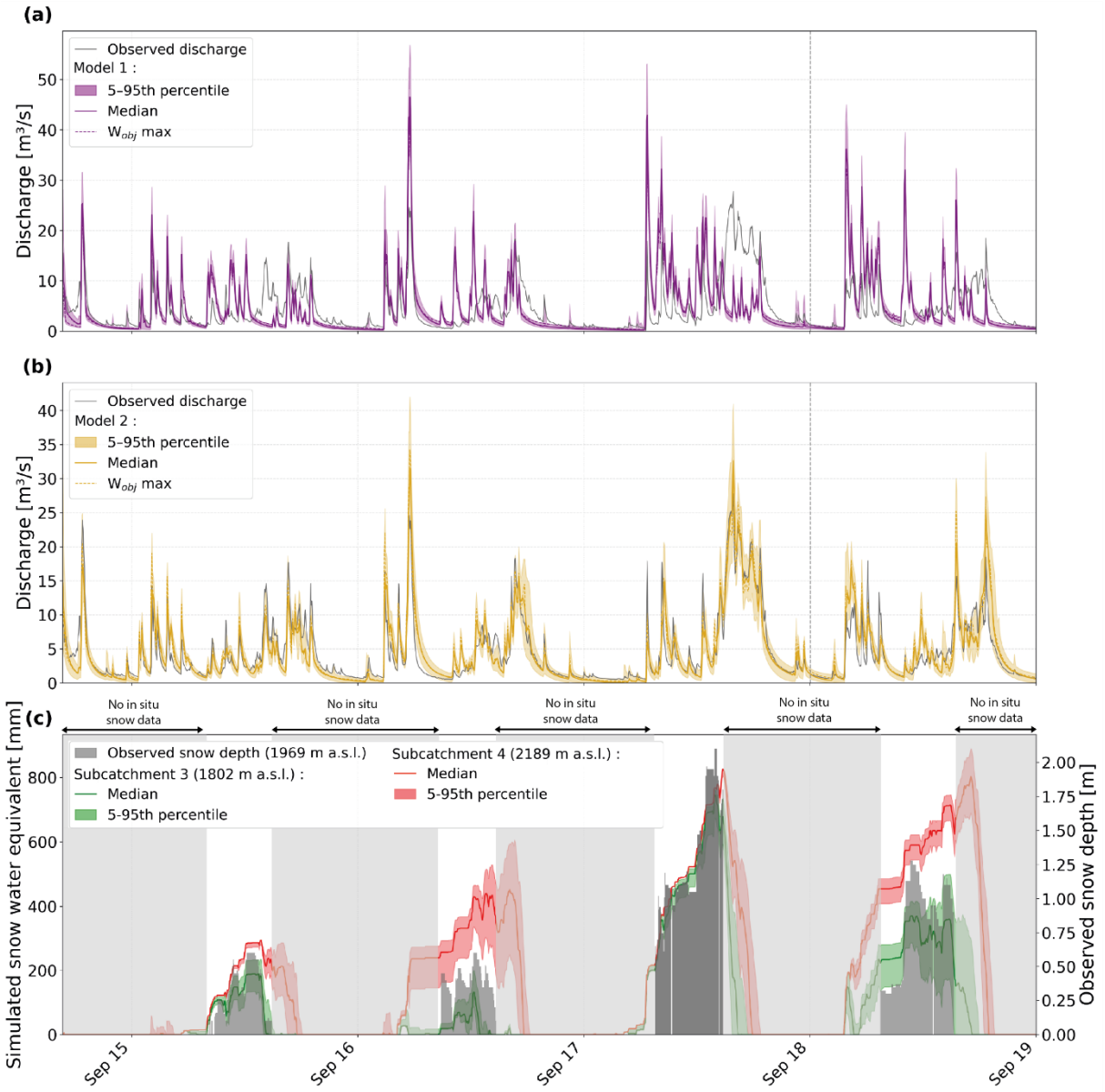


Figure 6: Observed and simulated discharge for the Dévoluy karst system (2015-2019) for (a) Model 1 and (b) Model 2. Shaded areas show the 5th-95th percentile of the 5000-parameter set, solid lines the ensemble median, and dashed lines the $W_{obj\_max}$ simulation. (c) Model 2 simulated snow water equivalent for subcatchments 3 and 4 (shaded: 5th-95th percentile) and observed snow depth at the Super-Dévoluy station (grey bars); gaps indicate periods with no available measurements.

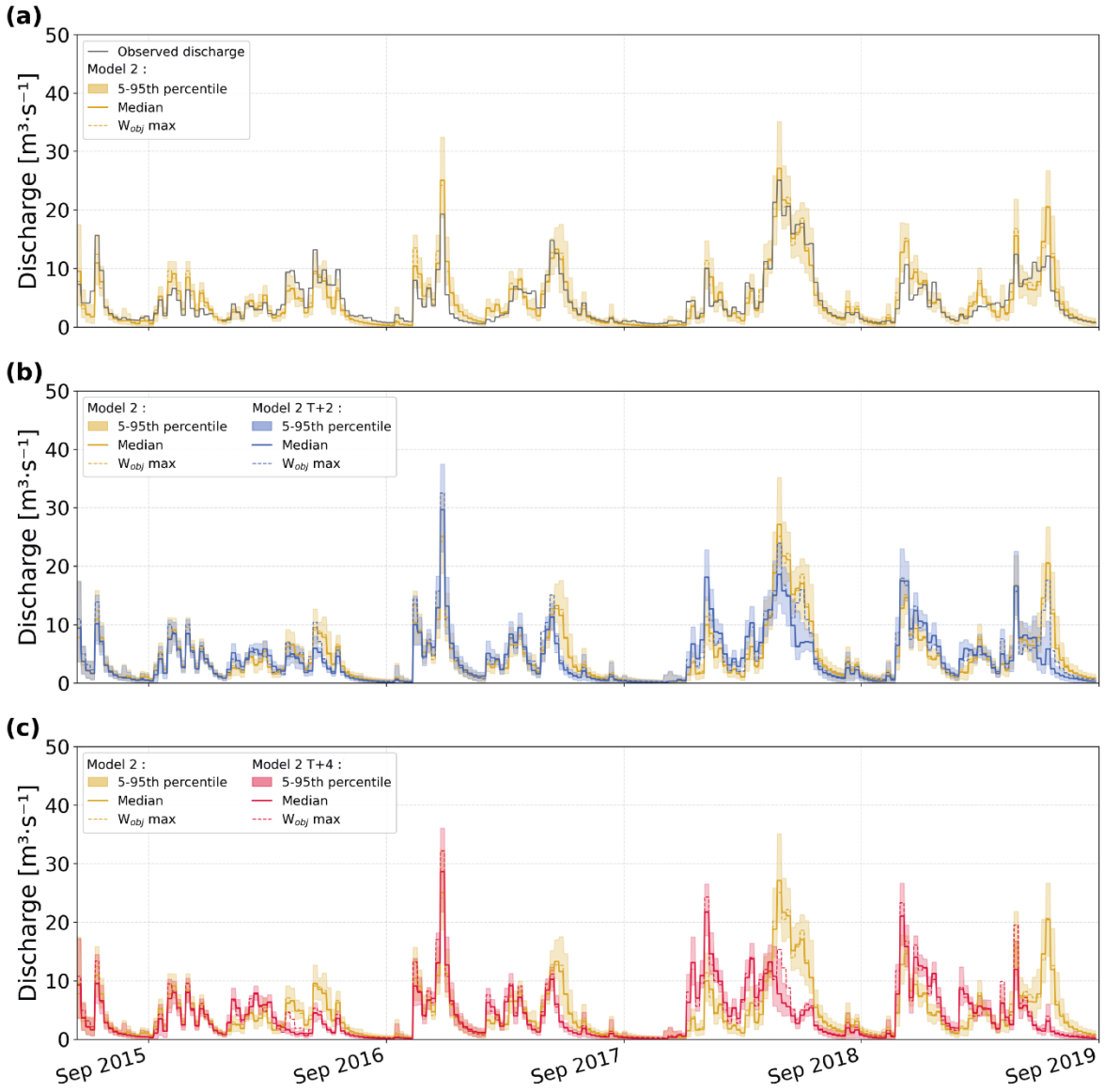


Figure 8**:** Seven-day average simulated and observed discharge for the Dévoluy catchment over the 2015–2019 period. (a) Model 2 and observed discharge. (b) Comparison of Model 2 and Model 2 T+2°C simulations. (c) Comparison of Model 2 and Model 2 T+4°C simulations. Shaded areas represent the 5th-95th percentile envelope of the 5000-parameter sets. Solid lines show the ensemble median and dashed lines the simulation with the parameter set corresponding to the maximum objective function value ($W_{obj\_max}$).

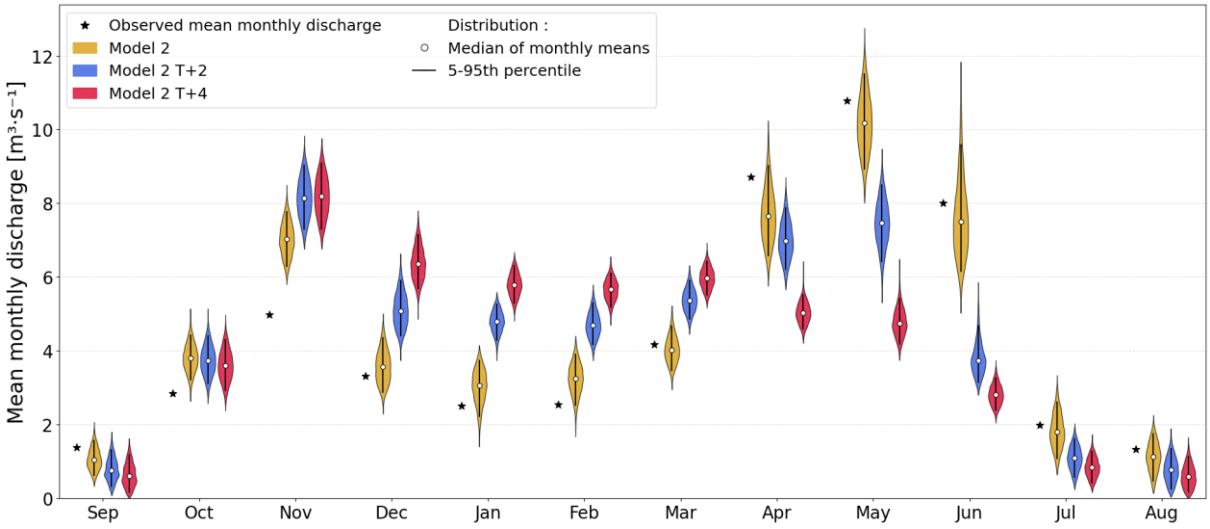


Figure 10: Mean monthly discharge distributions for the Dévoluy karst system over the 2015-2019 period, shown for each calendar month (September to August). For each month, the star symbol indicates the observed mean monthly discharge. Violin plots show the distribution of mean monthly discharge across the 5000-parameter ensemble for Model 2 (gold), Model 2 T+2°C (blue), and Model 2 T+4°C (red). The vertical bar and white dot indicate the 5th-95th percentile range and the median of the 5000-parameter ensemble.

References :

Çallı, S. S., Çallı, K. Ö., Tuğrul Yılmaz, M., and Çelik, M.: Contribution of the satellite-data driven snow routine to a karst hydrological model, J. Hydrol., 607, 127511, https://doi.org/10.1016/j.jhydrol.2022.127511, 2022.

Mazzilli, N., Guinot, V., Jourde, H., Lecoq, N., Labat, D., Arfib, B., Baudement, C., Danquigny, C., Dal Soglio, L., and Bertin, D.: KarstMod: A modelling platform for rainfall - discharge analysis and modelling dedicated to karst systems, Environ. Model. Softw., 122, 103927, https://doi.org/10.1016/j.envsoft.2017.03.015, 2019.

Sivelle, V., Labat, D., Mazzilli, N., Massei, N., and Jourde, H.: Dynamics of the Flow Exchanges between Matrix and Conduits in Karstified Watersheds at Multiple Temporal Scales, Water, 11, 569, https://doi.org/10.3390/w11030569, 2019.