6. It would be relevant to include the annual mean temperature and total precipitation in table 1 or in the study areas description section. These would help better understand the climate contrast among the different study sites.

Thank you for this suggestion, here are the mean annual temperature and precipitation for the 3 catchments (from ERA5-land reanalysis during the period 1991-2020) :

Catchment	Mean annual temperature (°C)	Mean annual precipitation (mm)
Albarine (FR)	10.0	1439
Genal (ES)	15.9	743
Lepsämänjoki (FI)	5.6	899

12. Line 157. Not clear if the period that you mention is for all study sites or whether it is for the calibration or validation step. Please clarify.

After the steps of calibration and validation, the JAMS/J2000 is used to reconstruct daily hydrological variables (discharge, groundwater contributions, evapotranspiration, snowmelt, soil saturation and groundwater saturation) for the period 2005-10-01 to 2022-04-30 for the 3 study sites.

14. Table 2. The Albarine and Lepsämänjoki sites have a continuous simulation period but the Genal has a gap between the calibration and validation period. Would it be relevant to explain why is that?

We encountered inconsistencies within the measured discharge data that raised concerns about the data accuracy. We analyzed the model performance in adjacent watersheds and found that the model performed well in the entire region. Hence, we concluded to leave out the periods, which are most certainly indicating potential errors. In an effort to maintain a collaborative and respectful relationship with the data providers, we aimed to handle the observed inconsistencies delicately and without causing undue concern. However, we acknowledge the validity of your point that mentioning these inconsistencies is crucial for transparency and accuracy in our research analysis. Therefore, we will mention this point in our final version.

15. The metrics in table 3 could also include a metric similar to your POD and FAR but for the hydrological model. For instance, set a low-flow threshold and indicate how many times the model succeeds simulating when the river is below the threshold (POD) or when the model is not simulating it accurately (FAR). This would also give information on how good the hydrological model is to simulate low-flow periods, which I think would add relevant information for your study. Is the hydrological model skillful simulating low-flow periods or is the RF algorithm making most of the work?

As you suggested, we analysed the performance of the JAMS/J2000 model to simulate discharges below a defined threshold at the gauging stations in the 3 catchments.
The chosen threshold is the 10th quantile of observed discharges on the total period (calibration and validation), keeping only the values for hydrological years with less than 10% missing values.

Then the performance metrics were computed as follows (the following metrics are different from the metrics used to analyse the performance of the RF model in the manuscript in response to a comment from referee RC2) :

Sensitivity = a / (a+c) (similar to Probability of Detection of drying events (POD) in the manuscript)
Specificity = b / (b+d) (similar to POD of flowing events)
Accuracy = (a+d) / (a+b+c+d) (probability of correctly simulating discharges below and over the 10th quantile)
True Skill Statistic (TSS) = Sensitivity + Specificity - 1

	Observed discharges	
Simulated discharges	Qobs < threshold	Qobs > threshold
Qsim < threshold	a	b
Qsim > threshold	c	d

a, b, c, and d represent the number of days for each of the 4 conditions during the calibration or validation period of the JAMS/J2000 model.

The Figure 1 shows the results of the analysis of the performance of JAMS/J2000 to simulate low flows.
For the validation period, Sensitivity is respectively equal to 0.84, 0.62 and 0.91 for the St-Rambert (Albarine), Lepsämänjoki, and Jubrique (Genal) gauging stations, which shows that the hydrological model simulates low-flow periods very well (for Saint-Rambert and Jubrique) and fairly well (for Lepsämänjoki).
However, for the St-Denis gauging station in the Albarine catchment, Sensitivity is equal to 0. This is due to the fact that the river is intermittent and sometimes completely dries at this station. For the St-Denis station the 10th quantile is equal to 0 m3/s, and as the JAMS/J2000 is not able to simulate complete drying, the model performance is poor.
These results show that the JAMS/J2000 hydrological model can provide correct simulations of the alternation between periods of low flow and medium-high flow to the RF model, but that the J2000-RF coupling is essential for correctly simulating the flow on intermittent river sections.

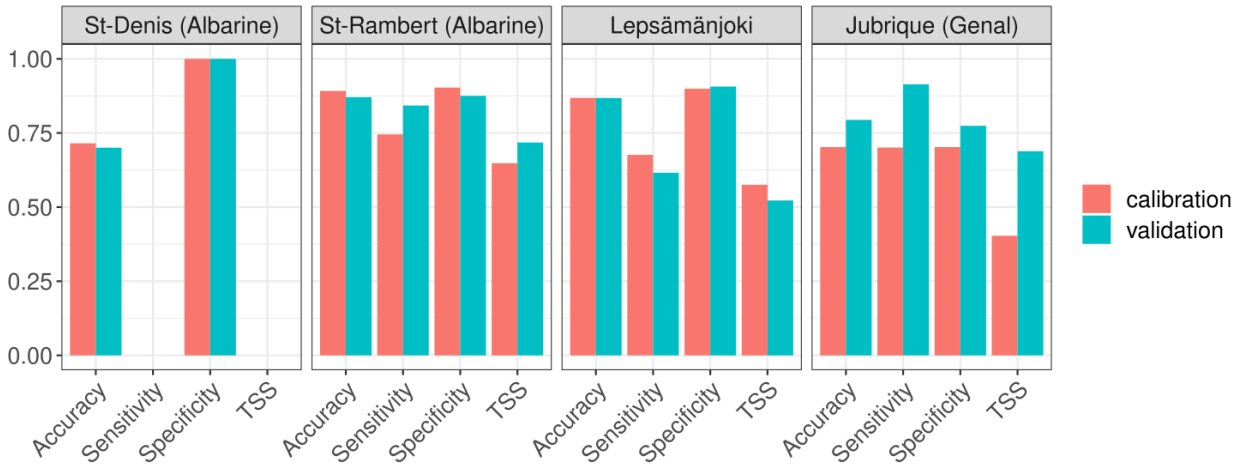

Figure A : Metrics for the prediction of low flows (< 10th quantile of observed discharges) with JAMS/J2000 model at the gauging stations for the calibration and validation periods.

17. Lines 167 and 278: Is it possible to infer why the uncertainty for the Genal catchment is higher? Is it that the catchment is very complex to simulate or some other issue?

We cannot say with certainty why the uncertainty is greater for the Genal catchment in Spain, but our hypothesis is that ther is a lack of observed flow state data to properly characterise drying patterns in the catchment, which is more complex than in the 2 other catchments (possible drying up

throughout the year and widespread drying up throughout the catchment). There is approximately the same amount of observed data in the spanish and finnish catchments, but the flow intermittence pattern in the finnish catchment is less complex (drying only during the summer month and only on small tributaries). Besides, the elevation and hence, slope range is much higher in Genal, which leads to a more heterogeneous pattern of drying, which also demands more observed data to increase robustness.

18. Line 279: Would it be relevant to also include some results for configuration 1 (perhaps in the supplement)? Just for comparison and to include something related to the uncertainty of the results.

Section « 3.5.1 Sensitivity to the size of the training sample » and more specifically Figure 11 already shows a comparison of the seasonal patterns of drying with configurations 0 and 1.

We agree that the sentence line 279 « The results presented in the next sections of this study were obtained with the configuration 0. » is confusing and will be removed. We will make it clearer that the spatial and seasonal patterns of flow intermittence as well as the covariates are first analysed with configuration 0, and that the sensitivity to the size of the training sample (configuration 1) or the type of observed data is analysed next.

We did, however, try to include some more results with configuration 1 to give a better idea of the uncertainty related to the size of the training sample. Figure B shows the importance of the covariates with configuration 1 in the 3 catchments. It is very similar to Figure 10 from the manuscript (with configuration 0) for the 5 most important covariates, which shows that for this study the importance of the covariates is not very sensitive to the size of the training sample, but rather to the quality of the covariates (cf section 3.5.3 Sensitivity to the geology data).
Figure B can be included in the supplementary material of the revised manuscript.

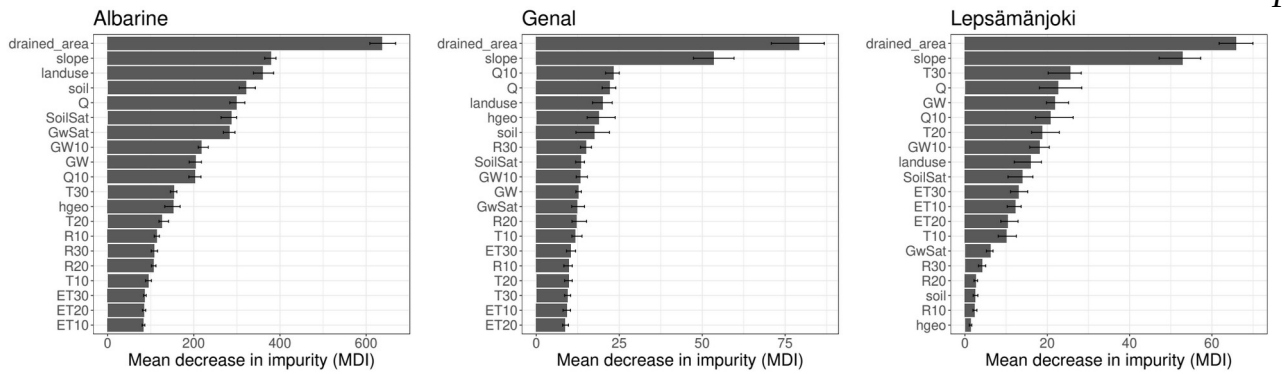

Figure B :Importance of the covariates in the RF models (mean decrease in impurity (Archer and Kimes, 2008)) for the 3 DRNs. Bars represent the mean MDI and the error bars the minimum and maximum

20. Lines 356 to 260. It is not clear if this paragraph only refers to the Albarine catchment. Please clarify.

Yes, lines 356 to 260 refer only to the Albarine catchment, as it is the only cachment for which we have another source of geological data. We will clarify this in the revised version of the manuscript.

22. Line 439. You could support this claim using previous expert elicitation studies that looked into expert perception uncertainty. For example: https://doi.org/10.5194/hess-26-5605-2022 or https://doi.org/10.1002/2015WR018461

Thank you for recommanding the 2 studies on expert elicitation. We propose adding the following sentences after line 439:

Expert elicitation in hydrology has already shown benefits, particularly when tangible data are missing (Ye et al, 2008, Warmink et al. 2011, Sebok et al. 2016, Sebok et al. 2022). These studies do show differences in the individual perceptions of the experts consulted, but by consulting a larger number of experts (in this study, only 1 or 2 experts were consulted per studied DRN) and by applying protocols similar to the ones proposed in these studies, the uncertainty linked to individual perception could be reduced, or at least quantified.