We warmly thank the Reviewers for their valuable comments. We provide below a detailed point-to-point reply to these comments. The changes included in the new version of the article are indicated in red.

1 Answer to Reviewer #1

GENERAL COMMENTS

The second revised version of the manuscript has satisfactorily addressed my remaining comments. I have only some minor comments of an editorial nature.

MINOR/EDITORIAL COMMENTS

(1) p. 8, line 23: "Section 3.1.1" instead of "Section 3.1"?

→ Done.

(2) p. 9, line 7: "far tail" instead of "very tail"

→ Done.

(3) p. 11, line 7: Notation "tr" could be confusing as it usually denotes the trace of a matrix.

→ Done: replaced by "$transf$"

(4) p. 15, legend of Fig. 4: Right and left descriptions should be reversed.

→ Done.

(5) p. 19, line 7 ("to illustrate the case of an heavy-tailed distribution"): Judging from the pattern of points on the return level plot (Fig. 7), not very strong evidence of a heavy tail.

→ We have modified the sentence: We illustrate the quality of the fit for the station Antraigues, located in the very foothills of the Massif Central slope (see Figure 1), which shows among that largest annual maxima (see Figure 2).

(6) p. 22, line 12: Insert "(i.e., a heavier tail)" after "more convex"

→ Done.

2 Answer to Reviewer #3

The authors have answered to all the questions of my previous review and have clarified some key-points of the manuscript.

Despite this, I still think that there is a significant open issue, that should be amended for the manuscript to be ready for publication. I still think that the advances that the proposed framework could provide to the scientific community does not emerge clearly from the manuscript: the objective framework can be quite useful from an operational point of view, but it ends up being a concatenation of standard approaches that consider a combination of standard indices for taking the final decision.

The authors state in answer 2 to my previous report that "they are not aware of any study comparing objectively the full procedure going from point measurements to mapped distribution". It could be true, but considering that in the literature there are plenty of methodologies for comparing different parent distribution and different mapping approaches, even separately (the authors list some of them in the manuscript (P2 LL 25-29), for stressing the relevance of their work the authors should do a better job in underling the advantages of using the kind of "unified objective framework" they propose. Is it faster compared to adopting the assessment of the marginal and of the mapping technique separately? Is it more efficient? More powerful in rank the performance of the different models?

This point is not explored at all in the manuscript. Section 4.1 and 4.2 are dedicated, as the authors state in answer 1 to my previous report, to describe the results of the application to the case study, analysing the performance of the tested models with the proposed indices. This can be used for showing the significance of the proposed indices, but can this be considered a validation of the framework on its own? As mentioned in *my previous revision, if a fair comparison with analogous techniques is not feasible, at least considering the results on a different basin could provide some sort of validation. Despite the basin is considered only for demonstration purposes, showing that the framework is able to provide the best distribution-mapping tool for any considered region, could be a good test for underling the potentialities of the framework in ranking the different considered models according to the input data.*

→ As suggested by the Reviewer, we have applied exactly the same framework to a much bigger catchment located in another climatic region, in order to show the generality of the framework. The chosen catchment is the Durance in Cadarache ($14,000$ km^2), located in the southern French Alps (see Figure 1). Note that altitude of the region ranges up to 4000 m.a.s.l whereas the highest peak of the Ardèche catchment are around 1500 m.a.s.l. We used 54 stations with data back to 1975. Figure 2 shows that the largest (resp. smallest) annual totals and annual maxima are about 2/3 the values of the Ardèche catchment (Figure 2 of the article). We use the same three weather types (WTs) as in the article. Based on Figure 3, we also define the season-at-risk as the three months of September, October and November (although we could have extended the season-at-risk to December and possibly January). We apply the cross-validation procedure proposed in the article to select both marginals and mapping models.

Figures 4, 6 and 8 show the same cross-validation criteria as in the article. Figures 5, 7 and 9 illustrate the case of the Coursegoules station which receives the largest annual totals (most south-eastern station in Figure 2). Figure 5 shows the benefit of considering subsampling into seasons and/or WTs. The gain in using both seasons and weather types is clearer for N_5 than for FF. Note that FF, N_5 and SPAN are lower than in the case of the Ardèche catchment, revealing a more reliable and robust fit of the tail. We select the model with two seasons and 3 WTs although a more parcimonious model could be to consider seasons only. For the Coursegoules station, the model with no WT (cases (1,1) and (2,1)) tend to underestimate the tail (see Figure 5). As for the Ardèche catchment, we select the Gamma marginal distribution but actually the exponential model is almost as good apart for N_5. For Coursegoules station the exponential and Gamma models are almost equivalent (see Figures 5 for (2,3) and 7). Finally, based on Figure 8, we select as in the article the bivariate thin plate spline interpolation using the smoothed altitude as covariate (tps2Z), which improves mainly the spatial robustness (TVD and KLD) compared to the kriging method, as illustrated Figure 9 for Coursegoules. Interestingly, we note that both the same marginal and mapping models are selected for the Ardèche and Durance catchment despite their different size and climatology. However we do not claim that these are the best models for any region. Marginal and model section is obviously region-dependent, so the results obtained here cannot be generalized to any region (even in France). Nevertheless we do claim that the proposed framework is general and can be applied to any region.

Although the results for the Cadarache catchment are interesting in that they prove the generality of framework, we have decided not to include them in the article because it is already quite long and because demonstrating the framework on the Ardèche catchment seems sufficient to us to illustrate how it works.

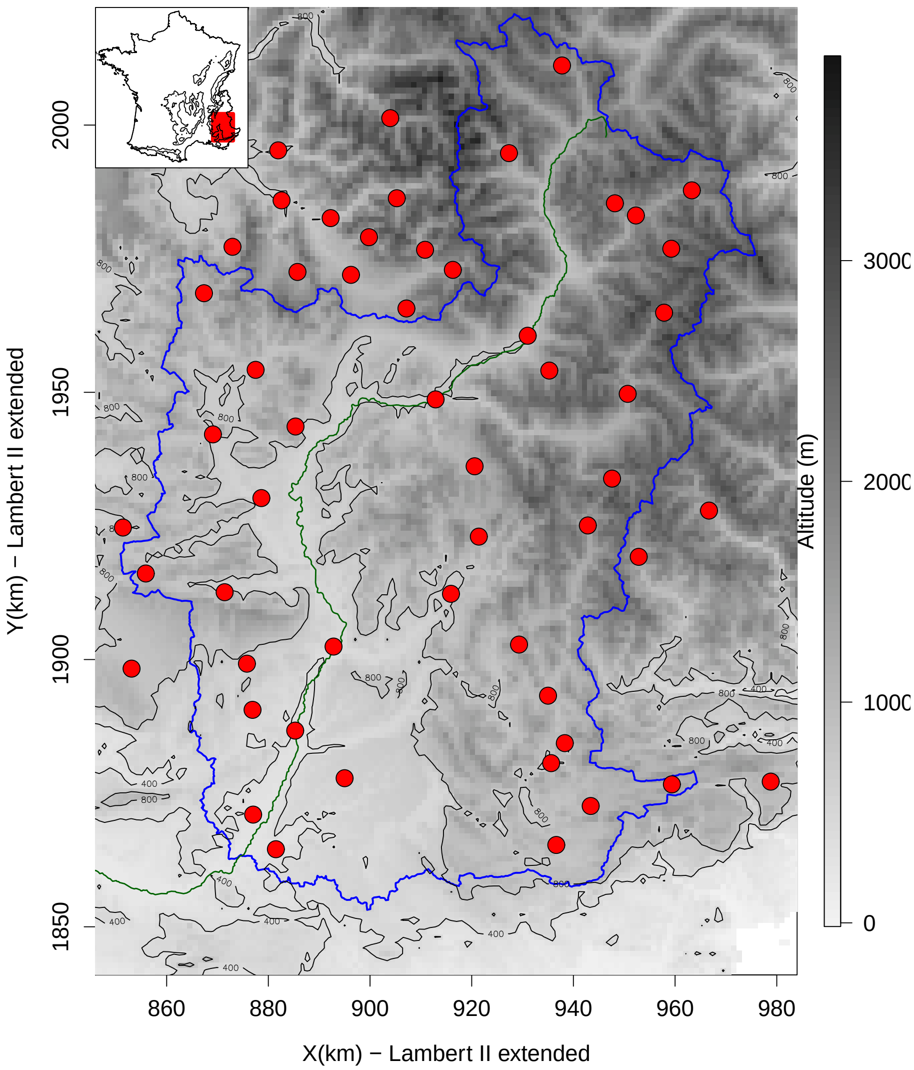

Figure 1: The Durance catchement. The red points show the location of the stations. The background shows the altitude in gray scale (1km raster cells). The top left insert shows a map of France with the studied region in red. The black lines are the 400 and 800 m.a.s.l. isolines.

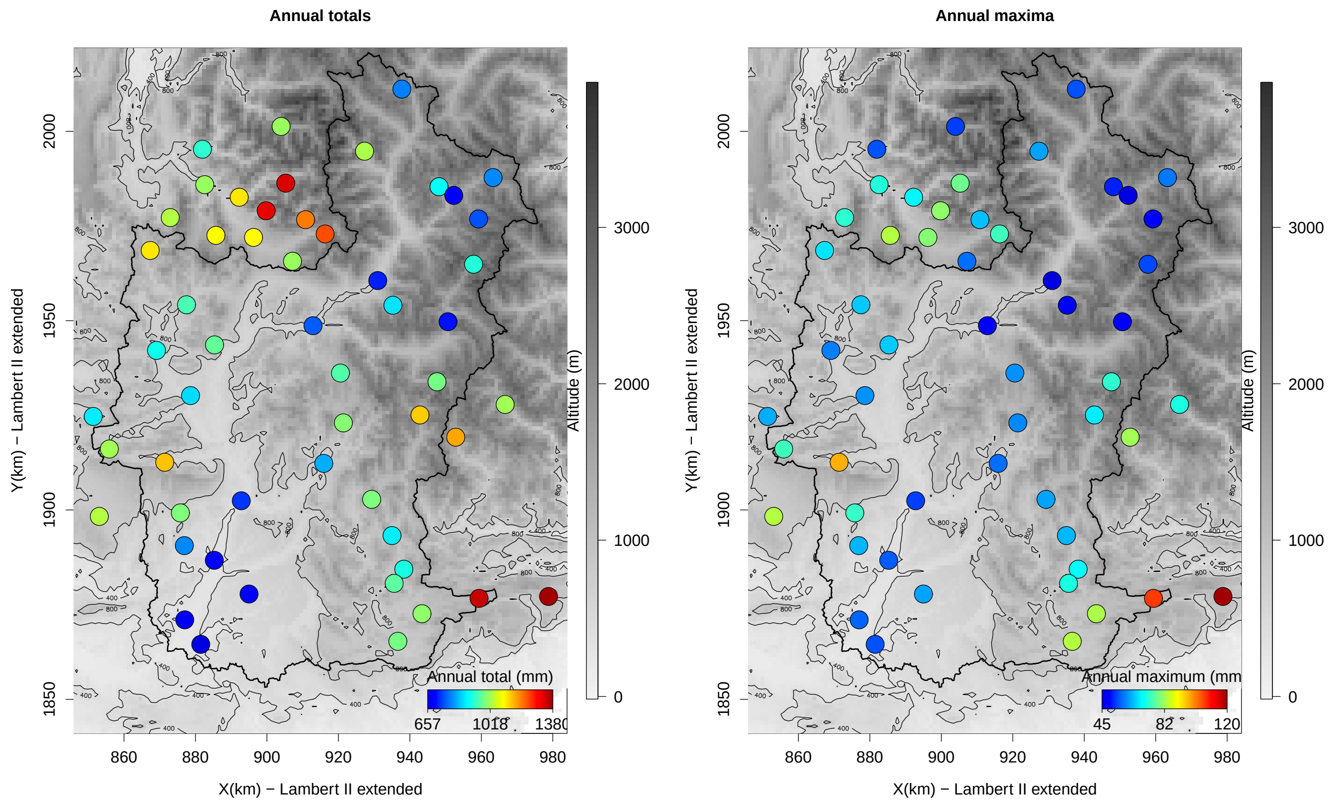

Figure 2: Left: Averages of annual totals (mm). Right: Averages of annual maximum daily rainfalls (mm).

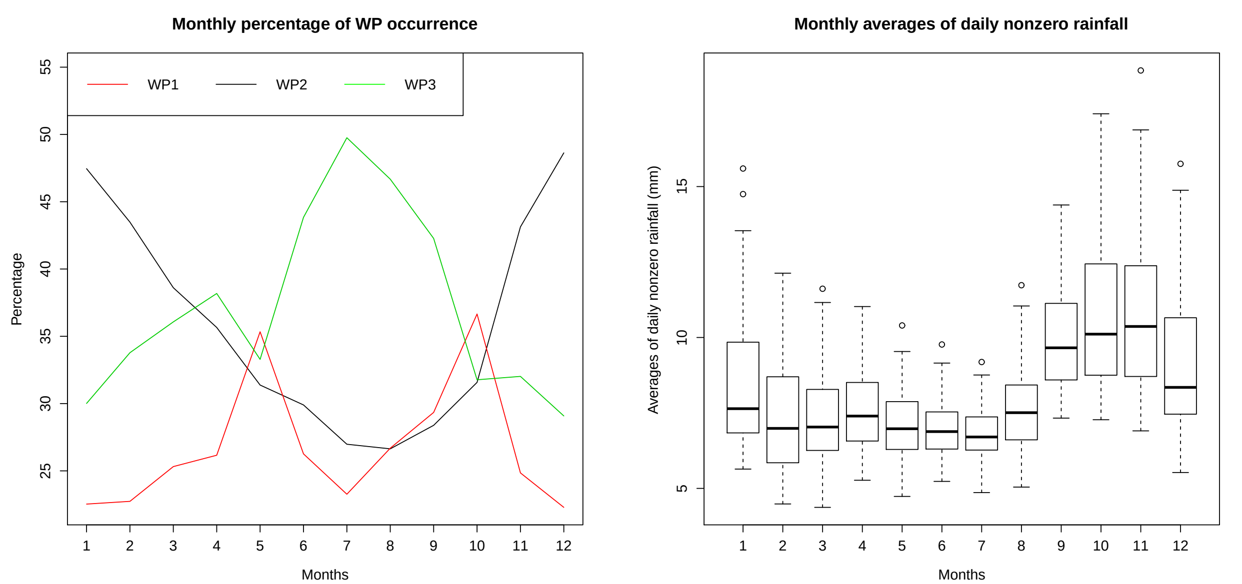

Figure 3: Left: Monthly percentage of occurrence of the three WPs. Right: Boxplot of the monthly averages of daily nonzero rainfall. Each boxplot contains 54 points (one point per station).

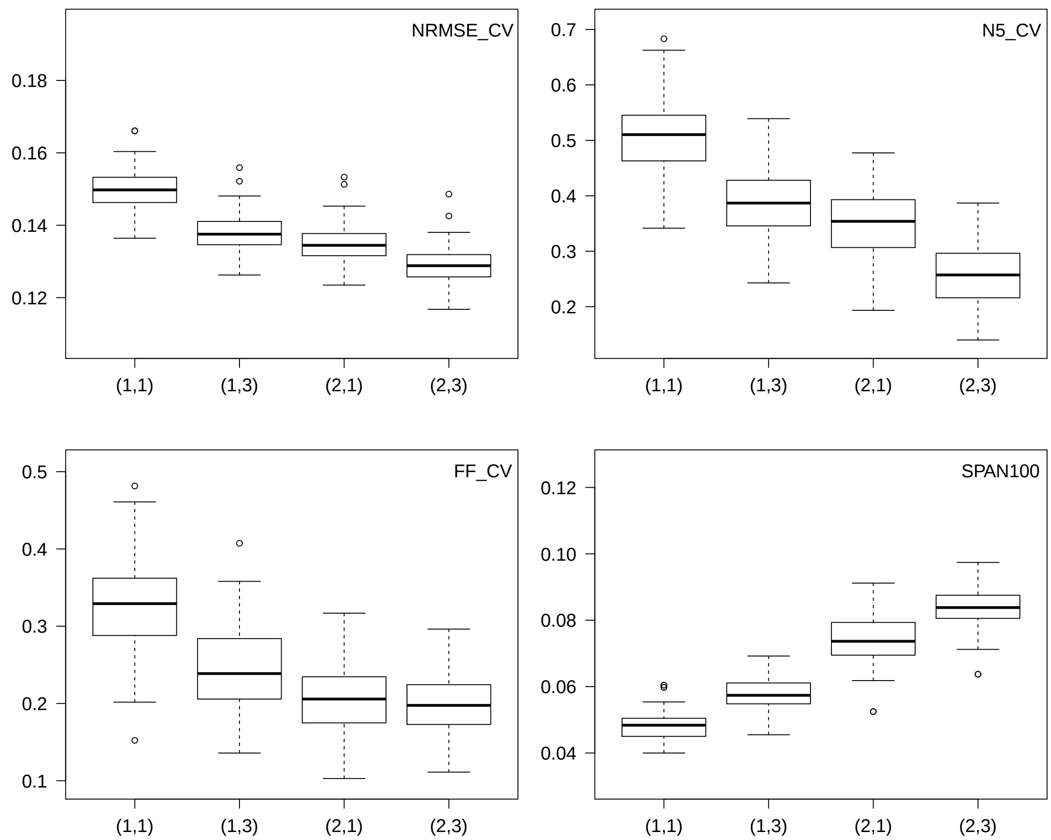

Figure 4: Scores of cross-validation when $G_{s,k}$ are Gamma distributions and the number of seasons and WP varies: $S \in \{1, 2\}$ and $K \in \{1, 3\}$. The values of (S, K) are indicated in the x-labels. Each boxplot contains 100 points.

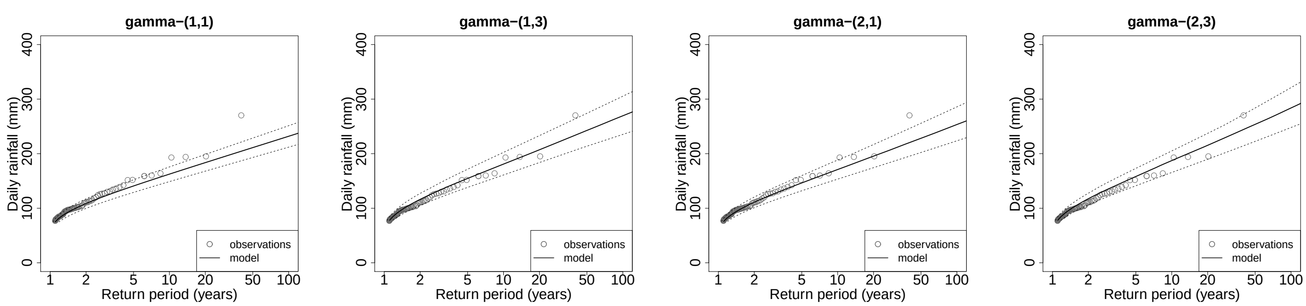

Figure 5: Case of Coursegoules when $G_{s,k}$ are Gamma distributions and the number of seasons and WP varies: $S \in \{1, 2\}$ and $K \in \{1, 3\}$. The values of (S, K) are indicated in the title. The dotted lines show the 95%-envelope of return level estimates over the 100 subsamples. The plain line shows the median estimates. The gray points show the full sample (35 years). Each estimation is based on half of these points.

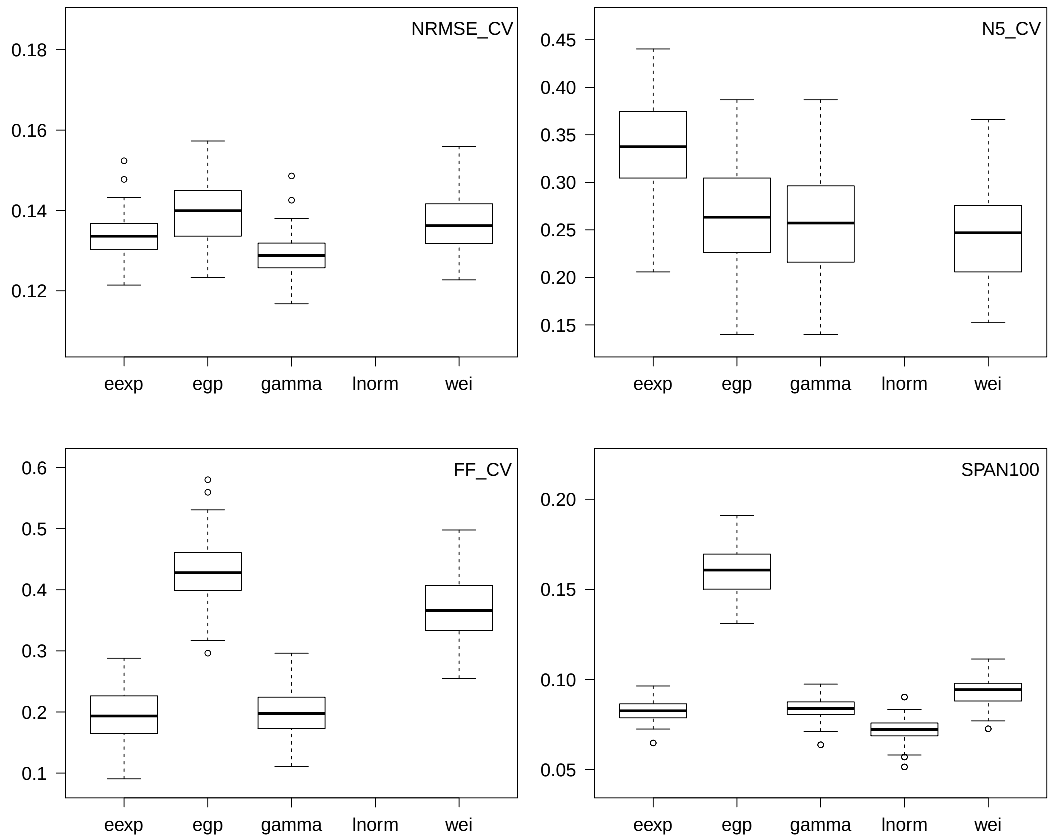

Figure 6: Scores of cross-validation when $G_{s,k}$ is either the extended exponential (eexp), extended Generalized Pareto (egp), Gamma (gamma), lognormal (lnorm) or Weibull (wei) distribution, with $S = 2$ and $K = 3$. Each boxplot contains 100 points. The boxplots of reliability scores in the lognormal case are missing because they lie far above the upper range of depicted values.

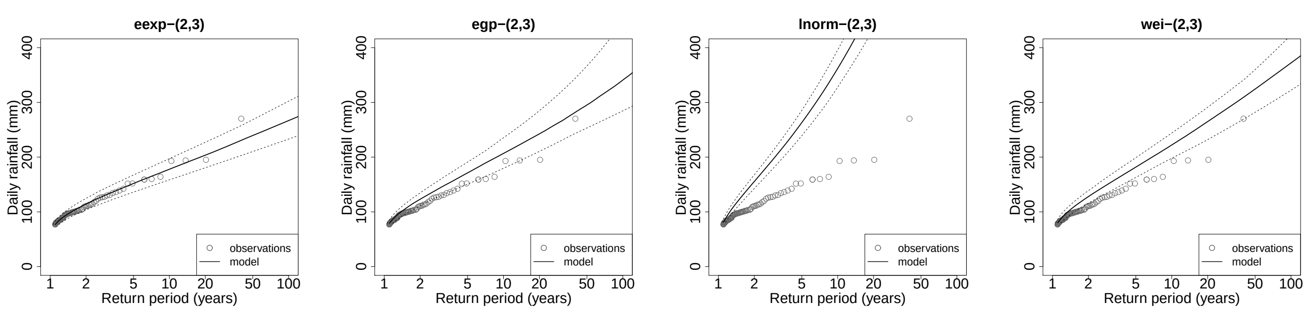

Figure 7: Case of Coursegoules when $G_{s,k}$ is either the extended exponential (eexp), extended Generalized Pareto (egp), lognormal (lnorm) or Weibull (wei) distribution, with $S = 2$ and $K = 3$. The dotted lines show the 95%-envelope of return level estimates over the 100 subsamples. The plain line shows the median estimates. The gray points show the full sample (35 years). Each estimation is based on half these points. Case of the Gamma distribution is shown in the right panel of Figure 5.

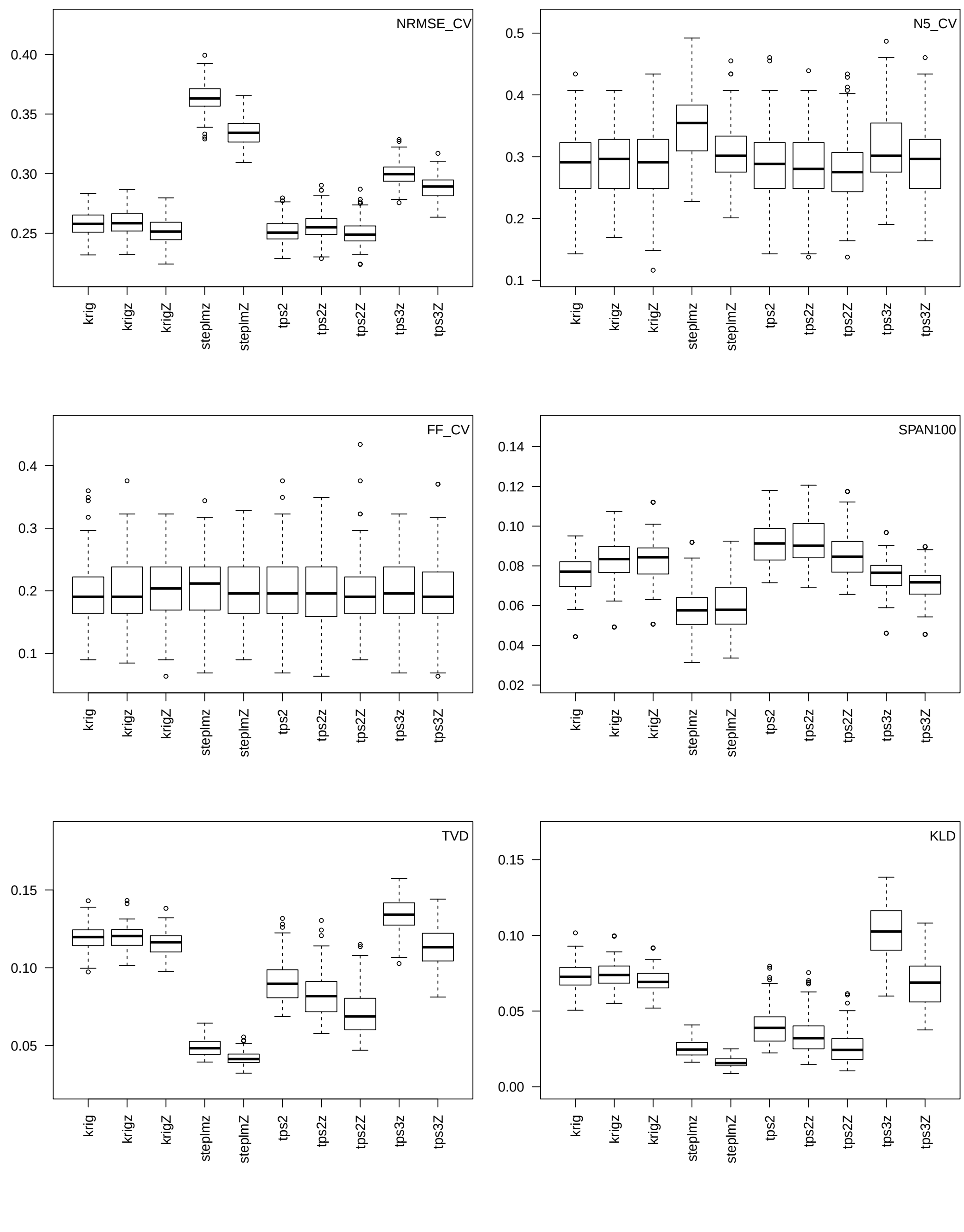

Figure 8: Scores of mapping when $G_{s,k}$ are Gamma distributions with $S = 2$ and $K = 3$ whose parameters are interpolated with the same mapping models as in the article. The two first rows show leave-one-out cross-validation scores. Each boxplot contains 200 points. The third row compares interpolations at a given station whether the data of this station are used or not in the interpolation. Each boxplot contains 100 points.

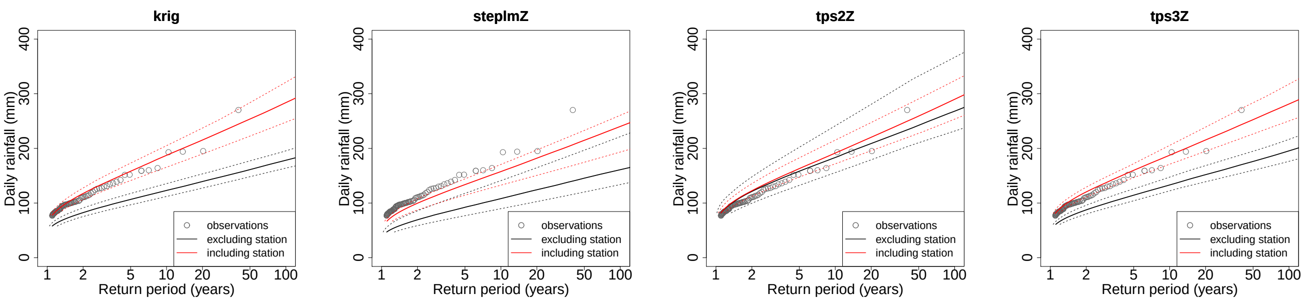

Figure 9: Case of Coursegoules when $G_{s,k}$ are Gamma distributions with $S = 2$ and $K = 3$ whose parameters are interpolated with either: kriging without extrenal drift (krig), stepwise linear model (steplmZ), bivariate thin plate spline with drift (tps2Z), or trivariate thin plate spline (tps3Z). The dotted lines show the 95%-envelope of return level estimates over the 100 subsamples. The plain line shows the median estimates. In black, each interpolation is based on half the data of the other stations, excluding the considered station. In red, interpolation is based on half the data of all the stations, including the considered station. The gray points show the full sample (40 years).