Response to Anonymous Referee #4

We thank Referee #4 for reading the manuscript carefully and providing thoughtful and constructive comments. The comments are noted with RC, our responses with AC, and the intended additions or changes in the manuscript are underlined.

This paper by Legrand et al. presents and discusses the results of a modelling chain that uses large-scale atmospheric information (ERA-20C reanalysis) to produce continuous daily river discharge at several stations of the Upper Rhône river catchment (~ 11000 km²). This modelling chain includes a semi-distributed hydrological model (GSM-SOCONT) and a downscaling model. Two downscaling models (statistical or dynamical) are actually used and tested. The results are analysed over several periods (1902-2009, 1961-2009, 1981-1983, 1920-1949, 1950-1979, 1980-2009) and according to various indicators (daily discharges, mean monthly and annual discharges, low and high flows).

This paper clearly corresponds to a huge amount of work. A lot of material and results are presented, the Figures are very rich.

We thank Referee #4 for his/her encouraging general comments.

The drawback of this very rich content is that it is not always easy for the reader to navigate and appreciate the results in the present form of the manuscript. My main remarks and recommendations to improve the paper are listed below :

Title and focus.

RC4.1 The paper is entitled "Simulating one century (1902-2009) of river discharges, low flow sequences and flood events of an alpine river from large-scale atmospheric information" which is a very general title and does not really correspond to the real focus of the paper. In my opinion, as stated by the authors themselves, the objective of the work presented here is to assess the ability of downscaling chains to simulate hydrologically relevant weather scenarios (p12, l. 282-284). This is consistent with the detailed description of the different steps of the downscaling models, and the presentation of the results as comparison to a reference that is not the observed discharge (but rather the discharge simulated by the hydrological model forced by reference observations). I therefore suggest to make this objective more clear in the Introduction, to change the title accordingly and to trim the parts of the text that do not contribute directly to this objective (a lot of details about the hydrological model description and set up could be placed in supplementary material for example).

> **AC4.1** Thank you for this comment. <u>In the introduction, we will explain more clearly our aim to better understand the crucial steps required to simulate hydrologically relevant weather scenarios.</u>
>
> <u>Note that a description of the novelty of our study will be included in the introduction, as this point was also highlighted by Referee #1 and Referee #3.</u> Please see our answer RC1.1 or RC3.1, which is the same.
>
> <u>As suggested, we propose to replace the title of the study with the following words:</u>
> "Hydrological implications of different downscaling approaches and bias correction of a global atmospheric reanalysis: case study of the Upper Rhône River (1902-2009)"

We decided to keep the part devoted to the hydrological model in the main article, as it contains some crucial and innovative steps, particularly those related to the presence of dams along the river, as well as the discussion linked to lapse rates.

Figures.

The Figures contain a lot of information, maybe too much compared to what the reader is able to see and to what is necessary to illustrate the author's point. They would gain a lot from a bit of trimming. As examples :

RC4.2 Fig 1 : the names of the gauging stations on the map are not necessary (they are also in Fig 2) and prevent the reading.

AC4.2 We prefer to keep the names of the gauging stations in Fig. 1, as this helps to identify ungauged sub-basins and to locate the gauging stations where hydrological assessments will be made later. For reasons of legibility, we will however use shorter names.

RC4.3 Fig 7 : is it really useful to present both MAT and DMAT ?

AC4.3 In Fig. 7b for MAT, it is rather easy to see the monthly biases obtained between observations and SCAMP simulations. However, the bias for the MAR simulation is difficult to identify due to the high seasonality of the raw variable. Fig. 7c is given to highlight this "small" bias, which nevertheless has strong implications on simulated hydrology. We therefore think it is important to keep both figures.

RC4.4 Figs 8 and 9 : these Figs are more illustrative, maybe present more synthetic results before (Fig 10) + it could be worth presenting a Table with a few synthetic values (bias, KGE) so that we can have a general idea of how the simulation chains perform compared to reference. Why presenting simulations with BC before without BC ? + is it really interesting to present the time series of lake level ? + add the names of the stations directly on the Figures

AC4.4 Presenting a Table with a few synthetic values**.**
Thank you for this suggestion. A Table with the NSE coefficients for the dynamical downscaling model (MAR/MAR-BC) and the statistical index CRPSS (Continuous Ranked Probability Skill Score) for the statistical downscaling model (SCAMP/SCAMP-BC) will be added to show the performance gain associated with bias correction.

Why presenting simulations with BC before without BC?
In the section devoted to the evaluation of SCAMP and MAR weather scenarios, we highlight the significant biases for both MAP and MAT variables. It was therefore logical for us to consider that hydrological simulations have to be performed with bias-corrected weather scenarios. This is why we present these results first. However, the presentation of the simulations with the raw scenarios follows. This allows us to show that the discrepancies obtained on discharges do not correspond to what we could have expected from the MAP biases, and this because of the MAT biases. In our opinion, this makes the demonstration of the importance of MAT biases more effective. This is the main reason why simulations without bias correction follow simulations with bias correction.

Is it really interesting to present the time series of lake level?
As the lake has a significant buffering effect on downstream hydrology, due to storage cycles, it was also important to check that the model is able to simulate the main features of storage variations.

Add the names of the stations directly on the Figures.
As suggested, the name of the gauging stations will be added on the figures.

RC4.5 Fig 11 : too many lines on Fig c) + Figs a) and b) don't work. It is not clear that there is a "column" for each period + the only thing we see is the comparison between SCAMP and SCAMP-BC for each period

AC4.5 Thank you for this comment. The columns for the 3 periods will be more clearly separated in Fig. 11a and b, and for greater clarity, Fig. 11c will be divided as follows:

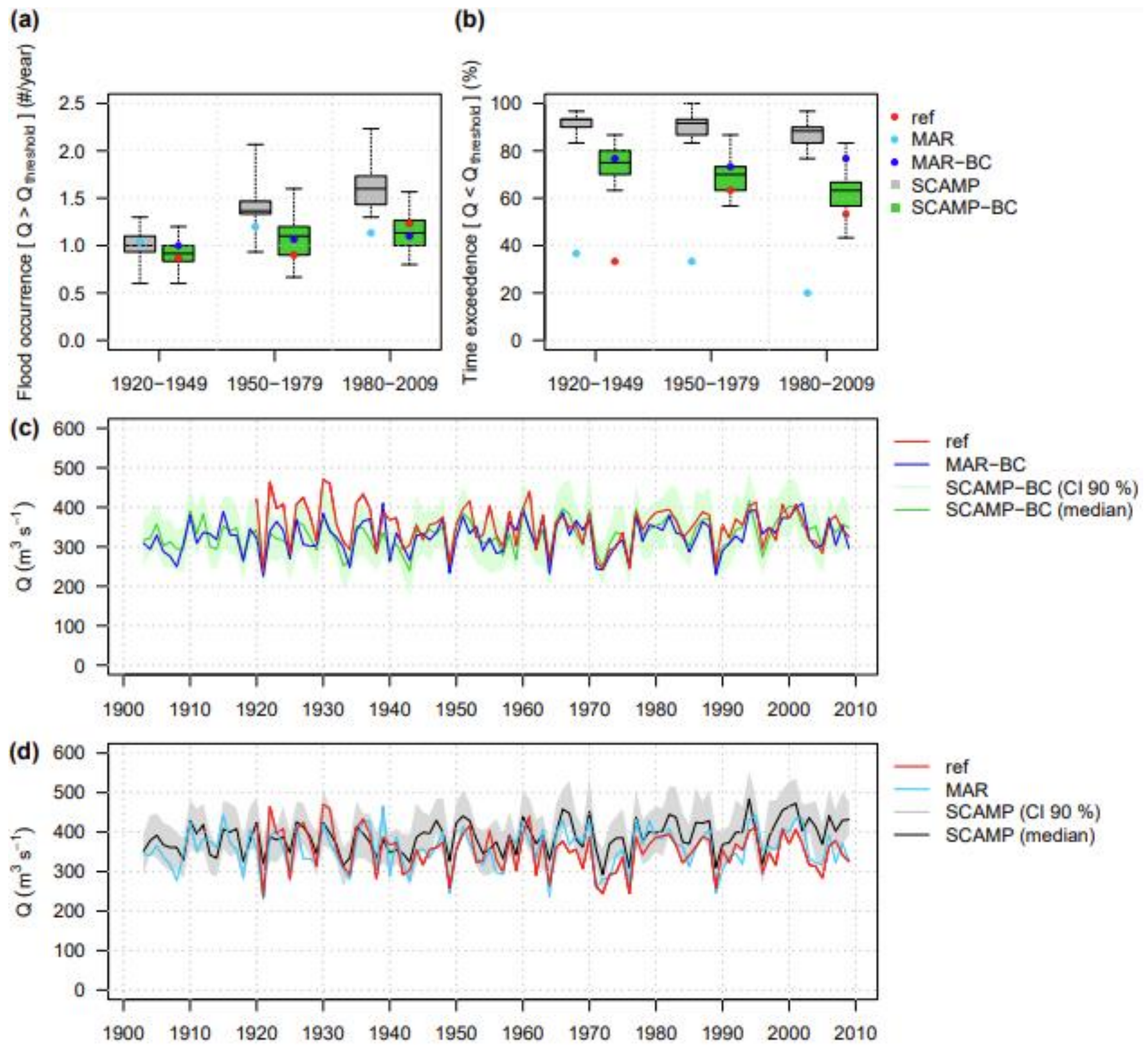

Figure 11. (a) Flood activity and (b) low flow activity at Rhône@Bognes for three 30-year sub-periods: 1920-1949, 1950-1979 and 1980-2009. (c) Mean annual discharges at Rhône@Bognes for the period 1902-2009 simulated with MAR-BC and SCAMP-BC. (d) The same for simulation with MAR and SCAMP. The grey and green bands represent the confidence interval at 90 % level. The median scenarios are indicated by the black and green solid lines. The "references" are observed discharges for the 1920-1960 period and simulated discharges from observed weather for the 1961-2009 period. MAR/MAR-BC: hydrological simulation forced by the raw/bias-corrected weather scenario produced with the dynamical downscaling model. SCAMP/SCAMP-BC: hydrological simulations forced by the raw/bias-corrected weather scenarios produced with the statistical downscaling model.

RC4.6 Fig 12 : Fig a) not legible. What am I supposed to see on Fig b) ? Why not removing MAR simulation (since lapse correction is only done to the MAR-BC simulation) ?

> **AC4.6** The size of the dots in Fig. 12a will be reduced for greater clarity. In Fig. 12a and b, note that the blue and red lines overlap. For greater clarity, this point will be added to the caption.
>
> In Fig. 12d, we show that the simulation with a classical bias correction of the Mean Areal Temperature (MAT) (bias correction of the mean for a reference altitude - black curve) leads to worse simulated discharges than those obtained without any bias correction (cyan curve). It is therefore useful to show the MAR simulation. The problem is clearly the bias in the temperature lapse rate. Simulations with double bias correction, i.e. 1) bias correction of MAT at a reference altitude, and 2) bias correction of temperature lapse rates, perform well (dark blue curve). Fig.12 a, b and c are different ways to highlight the importance of the bias for the temperature lapse rate. To the best of our knowledge, the discrepancy between the observed and simulated statistical distributions of the temperature lapse rate has never been recognized or illustrated before. In our opinion, it is therefore useful to highlight this issue here (Fig. 12b).

RC4.7 Fig 13 : Fig b) : why lines and crosses on this Fig for the different simulations ?

> **AC4.7** Thank you for this comment. We will also use lines for the MAP estimates with the two different precipitation lapse rates.

Acronyms and notations.

RC4.8 The paper uses many acronyms that are sometimes confusing and could be simplified. I particularly struggled with MAP / MAR / MAT (the name of the model can't be changed, but the names of the variables probably can), with basin-scale, RHHU, grid points, station (Fig 3 and accompanying text), see next remark. BC (Bias correction) is also not clear as it is used for different corrections (not the same for MAR and SCAMP, which I can understand, but also for temperature lapse-rate for MAR. About this last correction, if it is included in the results presented in section 5 (which is the case from what I can understand), why is it not presented along with the other corrections in section 4 ?

> **AC4.8** Thank you for this suggestion. For the notations, we will specify that we are referring to the dynamical downscaling model when we use the notation MAR. We will also specify the spatial scale at which the different MAP and MAT are considered in each section.
>
> Sect. 4.3 "Bias Correction" will be clarified and the equation for the lapse rate correction will be moved there, where we discuss the bias correction for MAP and for MAT.

Note that the principle of bias correction is the same for both models (MAR and SCAMP) and for all variables, i.e. MAP, MAT and temperature lapse rates. It is a quantile mapping bias correction. We simply estimate the correction function on different time series, depending on the model and variable considered.

Spatial resolution

RC4.9 I really struggled (and did not succeed) at understanding exactly what was done in terms of spatial aggregation / disaggregation for the weather scenarios in SCAMP. In step (1) how do to change from station (= point) observations to basin-scale resampled values (basin-scale = whole catchment or subcatchments?). Are the stations at the end of the process the same as the observation stations ? (I understand rather RHHU = subcatchment but later in the text it is still referred as neighbouring stations ad inputs to the hydrological model, see p 22 l 410-411). Similarly Fig 5 presents the numerical experimentation plan with the hydrological model forced by P and T at 7*7 km resolution. Please, make all this clearer !!

AC4.9 Thank you for this comment. There was indeed a mistake in Fig. 5, where disaggregated data are shown for MAR and SCAMP. SCAMP indeed produces precipitation and temperature time series at station level, and the stations at the end of the process are exactly the same as the observation stations.

<u>Fig. 5 will be corrected: the spatial resolution of P/T is “stations” for obs and SCAMP, and “7x7 km” for MAR.</u>

Sorry for this. We understand it was almost impossible to understand the spatial aggregation/disaggregation process. We hope it is now clear.

Time periods and indicators for the results.

RC4.10 As said before, there are a lot of results, presented for various times periods and various indicators that change from Figure to Figure. This does not make the reader’s task easy. I would be nice, along with the experimental setup, to define in a Table the various indicators and time periods used and explain why they were selected.

AC4.10 Thank you for the suggestion. <u>We will add a table to the Experimental setup section, with the indicators, the assessment objectives, the time periods and the areas considered for the evaluations.</u>

RC4.11 If I am not mistaken, the results of the calibration of the hydrological model for the reference simulation are not presented. It would be nice to add a few elements about this, just to make sure that the reference model is not completely off the track.

AC4.11 The calibration efficiency is already mentioned in Supplementary Materials (Table S1):

- For all sub-basins for which a classical calibration was carried out, we give the NSE coefficient calculated from observed and simulated time series of discharge.
- For all sub-basins for which a signature-based calibration was carried out, we give the NSE coefficient calculated from observed and simulated regimes and the Kolmogorov-

Smirnov coefficient calculated from observed and simulated distributions of annual discharge maxima.

Reference MAP and MAT.

RC4.12 I am not a specialist of estimation of weather variables in montainous areas but I am surprised by the methodology used here, which consists of Thiessen polygons with a density of stations that is not so high (62 raingauges for ~ 11000 km², even less temperature stations). There are several other methods for the estimation of areal P and T, from very simple such as inverse-distance weighting to more complicated (kriging). What is the reason of choosing such a simplistic approach ? How confident can we be with these “reference” values ? This should at least be discussed thoroughly.

AC4.12 As mentioned previously, spatial variations in weather in a mountainous environment can be obviously significant, often due to the effect of topography. A satisfactory estimate of MAT and MAP is however impossible to achieve from stations only, even in this Upper Rhône River configuration where the network of precipitation and temperature stations is denser than in many basins worldwide. In our experience, it is even rather inextricable.

The inverse distance is of course interesting, but has other limitations: the distance to be considered in a mountainous environment is not trivial at all. Two stations distant by 5 km in two different neighboring valleys are probably much more distant than 2 stations distant by 10 km in the same valley. A topographical distance has been proposed in some papers, but the choice of weight linked to topography has to be estimated and is always based on different assumptions and choices that are rarely justified (at least from a physical point of view).

Using an inverse distance approach to estimate MAP is not necessarily straightforward either. It would be necessary to estimate the weights of all stations for all (grid) points in the spatial unit considered. In most cases, the weights are estimated from the distances to the centroid of the spatial unit, which is also an important and not really satisfactory assumption. Let's consider 2 basins A and B of the same shape, very elongated (this is often the case in the Upper Rhône River catchment), with the same surface area and centroid. The only difference between the two would be their orientation: basin A is oriented in a West-East direction, basin B in a North-South direction. With an inverse distance approach, all weather stations would have exactly the same weights for both basins, which would obviously not be the case in reality.

Kriging would also be interesting. We could try Kriging with external drift, where the drift is linked to the topography. A number of works have been presented on this subject. However, they generally apply to interannual mean variables, i.e. mean annual precipitation, mean annual temperature. At these aggregated scales, there is indeed a clear relationship between meteorological conditions and altitude. However, this relationship is no longer valid for each individual time step of the period to be considered. As mentioned in the article, the elevation-temperature relationship (if any) varies from day to day. Kriging is much more computationally demanding than Thiessen polygons. It also requires the calibration of a functional relationship (the variogram model). The quality of this model must be verified. It must be verified for each time step, which would be really difficult to do.

For precipitation, as mentioned in the article, the nonexistent precipitation-elevation relationship at high resolution and at the scale of precipitation events is another important limitation.

For these reasons, and for the sake of simplicity, we have chosen to work with the Thiessen method. Note that estimates of MAP and MAT are obtained from a rather high number of stations (5 to 10 for each of the 18 sub-basins for example for precipitation). We therefore expect not to produce very poor estimates of MAP and MAT.

Note also that this choice should not influence the main results of our work, notably the critical issue of bias in downscaling models, especially that of the temperature lapse rate for dynamical downscaling, and the critical issue of the precipitation lapse rate at the event scale.

We agree that this point should ideally be discussed in depth. However, the article is rather long and covers many other issues. We will just mention it briefly.

RC4.13 p 24 the authors write "no precipitation-elevation relationship was considered" although p 6: "and a regional and time-varying elevation-temperature relationship".

AC4.13 The issue of elevation dependency is different for precipitation and temperature.

For temperatures, the dependency is very strong and robust for most time steps, and is accounted for in (likely) all hydrological models (for mountainous areas at least, where snow is important) with a linear elevation-temperature relationship. The slope (the lapse rate) of this relationship can vary from one time to another. We have accounted for this temporal variability, estimated from station observations (see Sect. 4.1.1).

For precipitation, as discussed in Sect. 6.3, there is no clear precipitation-elevation relationship at the event scale, and the inclusion of the mean relationship observed for a long time period is problematic for flood simulation. This is why our simulations do not take into account a precipitation lapse rate. To the best of our knowledge, the precipitation-elevation issue was not described in previous works using hydrological models. However, it is potentially critical. We hope that our work will lead scientists to pay more attention to this issue in future studies.

Floods.

RC4.14 I have a few comments on the "flood" part of the paper. First, I do not really agree with the use of the term "flood" given what is presented here. In 5.3 (Fig 10) what is done is picking the maximum daily flow for each year, which does not necessarily correspond to a flood. Similarly, flood events can be expected to last several days given the size of the catchment so I don't think that the max daily discharge can be called a flood event.

AC4.14 Thank you for this comment. We agree with the reviewer's comment: we analyze the annual maxima of daily flows and they do not always correspond to floods. We will specify in the article that, for this part, we use the annual discharge maxima as flood proxy indicators and replace the term "Flood events" with "Annual discharge maxima" in Fig. 10.

RC4.15 In 5.4 (Fig 11), I suppose that the "flood activity" is defined as the number of days over the threshold, which again does not necessarily correspond to single flood events. The way of calculating the thresholds should also be precised (p 21, l 373-374 : are they defined through a flood frequency analysis as the value of the flood return period 1 year, or from the flow duration curve ? Therefore I would be more cautious with the term "flood" and would preferrably use terms like high flow indicators or maybe flood proxy indicators.

> **AC4.15** For flood activity, the threshold was calculated by considering the 90 largest flood events over a 90-year period. If the threshold was exceeded on several consecutive days, only the date of the maximum annual discharge over a 5-day window was retained. In this analysis, several flood events could be identified in the same year. Here we come close to the definition of "flood" referred to by the reviewer.
>
> For low flow activity, the threshold was calculated by considering the median of the 90 minimum annual reference flows at a monthly time step. Here again, several low flow sequences could be identified in a single year.
>
> <u>These points will be clarified in the text.</u>

RC4.16 p 26, l 483-488 : the unrealistic simulated discharges obtained with altitude-elevation correction can also be due to the calibration of the hydrological model, considering in particular that the model was "high flow calibrated" (4.1.2). A rigorous way to test that would be to re-calibrate the hydrological model with the new corrections added to the observations before testing them on the downscaling models.

> **AC4.16** Thank you for this comment. A recalibration would obviously be interesting, but based on previous analyses, we argue that it would not allow to fix this issue in a relevant way.
>
> As mentioned in the article, if an elevation-precipitation relationship can be identified from aggregated (e.g. annual) precipitation data, this is not the case at the scale of individual events. This is illustrated by the daily rainfall amounts observed at the different stations in the catchment area for the 4 largest floods recorded during the period. As shown in Fig. A below, there is no dependency on elevation. For event 2, precipitations at the highest elevation are even lower than in the lowlands.
>
> As shown in Fig. B below, a precipitation lapse rate of 10%/100m significantly increases the precipitation amounts, even for the heaviest precipitation events. On average, the annual MAP maxima at the Upper Rhône River basin scale are increased by 20%. Due to the high non-linearity of the production processes, this 20% increase leads to a much higher increase in peak flows. A recalibration of the model would reduce the "productivity" of the catchments, i.e. it would increase the storage capacity of the soil reservoir. The filling rate of this reservoir determines the runoff coefficient of the catchment for that period, and the lower the capacity, the easier it is to fill the reservoir. To avoid huge and unrealistic floods in a +10%/100m lapse rate configuration, a much higher soil capacity is therefore required. Previous analyses show, however, that this can be detrimental for smaller (and more numerous) rainfall-runoff events, leading to a number of minor floods being underestimated.

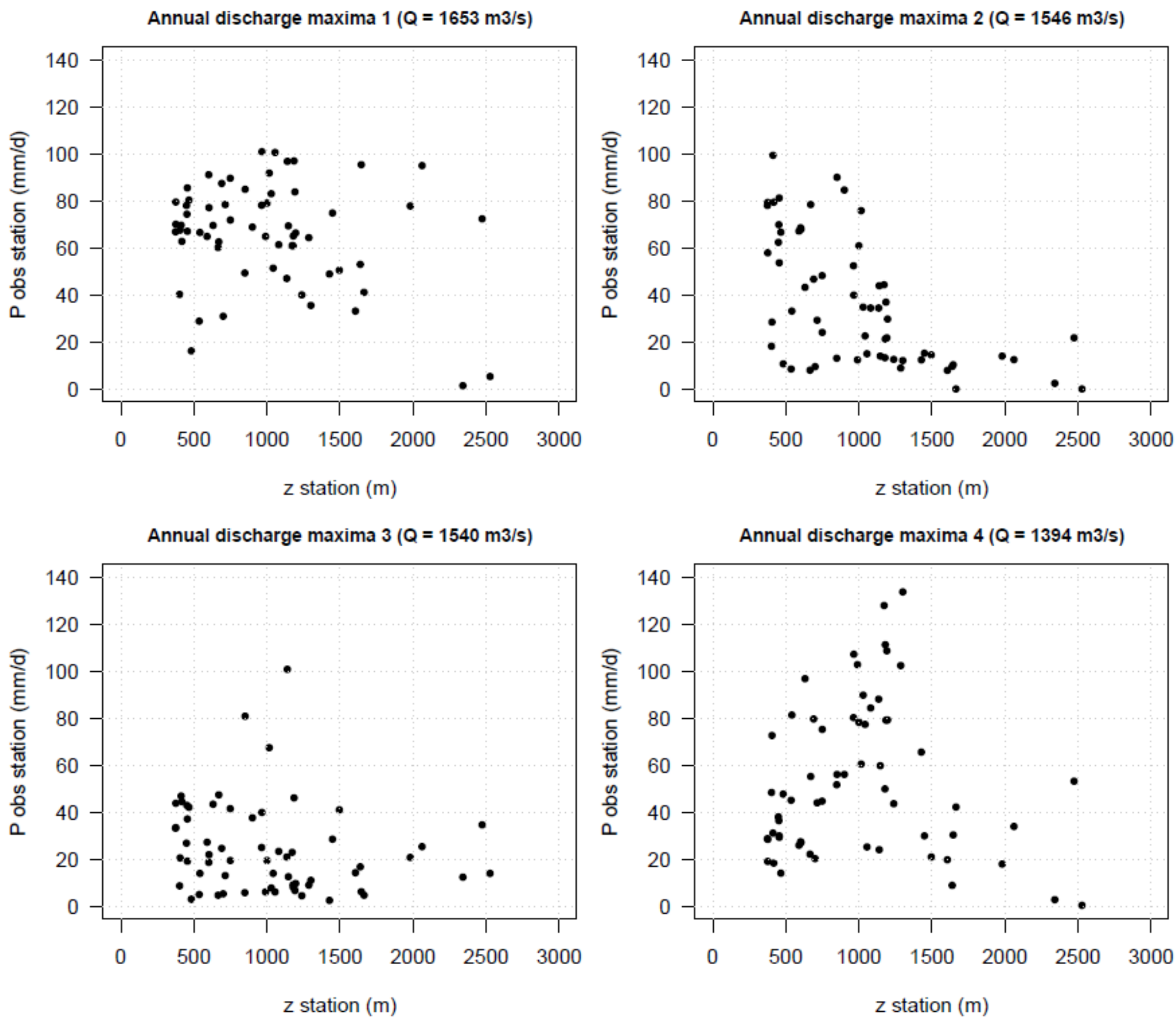

Figure A.

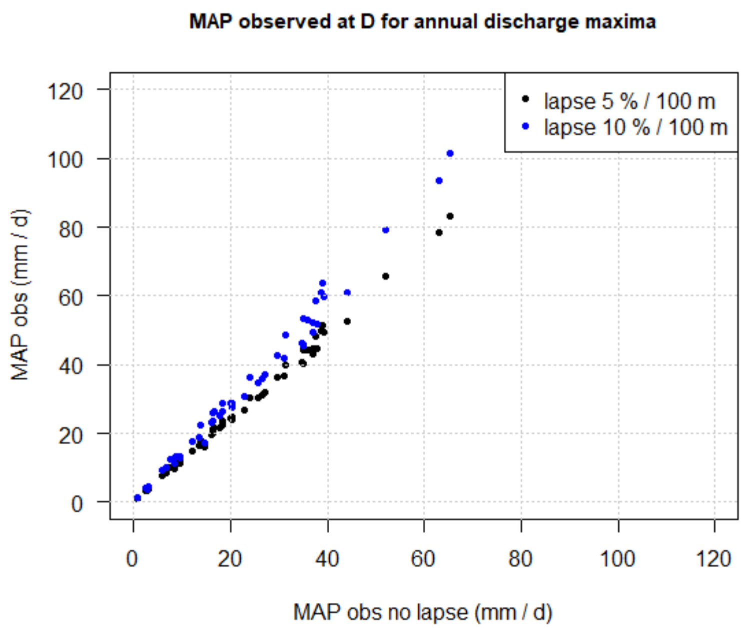

Figure B.