Response to Reviewers – HESS-2021-156

Oscar M. Baez-Villanueva, Mauricio Zambrano-Bigiarini, Pablo A. Mendoza, Ian McNamara, Hylke E. Beck, Joschka Thurner, Alexandra Nauditt, Lars Ribbe, Nguyen Xuan Thinh

Dear Jim Freer (Associate Editor), Juraj Parajka, Anonymous Reviewer 2, and Elena Toth,

We hereby provide the responses to the reviewer comments for our article *"On the selection of precipitation products for the regionalisation of hydrological model parameters"*. The revised manuscript is attached, including a tracked-changes version. The major changes to the manuscript are as follows:

1. Inclusion of a more detailed comparison of P products, considering differences in seasonal distribution, frequency, dry and wet spells, and daily extremes for the entire period as well as for the calibration, Verification 1, and Verification 2 periods;
2. Evaluation of the independent calibration and two verification periods according to hydrological regime;
3. Separation of the regionalisation results into the same time periods corresponding to the calibration (2000–2014), Verification 1 (1990–1999), and Verification 2 (2015–2018), thus improving the comparability of the results.
4. The inclusion of Section 5.2, which explains how the calibration of TUWmodel compensated for differences between the P products;
5. Evaluation (Section 4.4) and discussion (Section 5.5) about the influence of the number of donor catchments used in feature similarity; and
6. Removal of one of the two analysed objective functions (aggregated objective function; AOF) from the manuscript, and a greater focus on the analysis of the results using the remaining objective function (KGE').

We would like to thank all reviewers for their constructive comments and suggestions. We firmly believe that the quality of the manuscript has increased substantially by implementing these suggestions. Please find in the following pages our detailed responses to all your comments, including the modifications we have made to the manuscript. We hope that all the points you raised have been clarified and thank you again for your time and effort.

Sincerely,

Oscar Manuel Baez-Villanueva and M. Zambrano-Bigiarini, on behalf of all authors

Reviewer comments

Juraj Parajka (JP)

JP–General comments: This study examines the impact of four gridded daily precipitation products on daily runoff simulations in ungauged sites. The evaluation is tested using a large sample of near-natural catchments in Chile. The regionalisation of model parameters includes three methods (spatial proximity, similarity and regression). The results are evaluated by various runoff efficiency criteria and are compared over diverse hydrological regimes across the study region. The results show any noticeable impact of spatial resolution of precipitation product on runoff model efficiency. Also, the precipitation products corrected with daily gauge observations did not necessarily translate into improved hydrological model performance. The precipitation products with the best performance during calibration and/or verification did not necessarily provide the best simulations in ungauged sites. The best regionalisation approach is the similarity, regardless of the choice of gridded precipitation product or the calibration criteria. The results indicate that the performance of regionalisation depends on the hydrological regime.

The study presents an interesting analysis that adds some new understanding of the impact of using different precipitation inputs to predict daily runoff hydrographs in data-sparse and ungauged catchments. The manuscript has a good structure and reads well. I like the comparative focus of the analysis. The use of a large sample of catchments in very diverse climate and runoff generation conditions brings interesting and new findings. I have only a few remarks or suggestions which might be considered for extending some of the results and supporting interpretations made. These include:

We thank the reviewer for these kind words and comments, which have motivated us to improve the quality of the manuscript considerably. The specific comments are addressed below.

JPC1: The main focus is to evaluate the impact of selected precipitation products on regionalisation model performance. Still, the description and demonstration of the similarity and differences between the products and eventually at site measurements are very brief. Comparing of the long-term averages gives some impression about the consistency between the products, but it will be very interesting to see (and evaluate) some more detailed analysis of the differences between these products, such as the difference in seasonal distribution, intermittency/frequency of rainfall, days without rain, daily extremes etc. Can such additional differences/similarities explain more the differences between the results (i.e. calibration/validation/regionalisation performance, methods, time periods)?

Thank you for this comment. We agree that, given the comparative focus of our analysis, including a more detailed comparison between the P products ensures that the manuscript is more complete. Such a comparison provides a better understanding of how the differences between these products are translated into differences in performance during calibration, verification, and regionalisation. Due to the large variety of climates, we considered that the five macrocliamtic zones described in

Zambrano-Bigiarini et al. (2017) were the most appropriate for representing the climatic conditions of Chile. These zones have been added to Figure 1 of the manuscript and described in L84–87:

"A large variety of climates are present across the country, with the macroclimatic zones transitioning from the (hyper)arid and semi-arid climates in the Far North (17.50–26.00°S) and Near North (26.00–32.18°S), through temperate climates in Central Chile (32.18–36.40°S), to more humid and polar climates in the South (36.40–43.70°S) and Far South (43.70–56.00°S)."

Next, we derived the mean monthly areal P over the catchments located within each macroclimatic zone for the entire evaluation period and added this to Figure 2 of the manuscript (see also ETC3). This figure (replicated below as Figure R1) now shows both the spatial patterns of long-term mean annual P as well as mean monthly P values for all the catchments within each macroclimatic zone. To differentiate between the near-normal and dry conditions (suggested in ETC3), we recreated this figure for the calibration (2000–2014, near-normal, Figure R2), Verification 1 (1900–1999, near-normal, Figure R3), and Verification 2 (2015–2018, dry, Figure R4) periods. These figures have been added to the supplement of the revised manuscript. The following text has been added to describe Figure 2 (L154–171):

"Figure 2a shows the spatial distribution of mean annual P for all products over 1990–2018, while Figure 2b shows boxplots of the mean monthly P averaged over catchments located within each macroclimatic zone. All P products show relatively similar patterns of spatial variability across continental Chile; however, there are substantial differences in their total P amounts. In general, P increases from the (hyper-arid) Far North to the South, and decreases again in the Far South. P also increases from the west coast towards the Andes Mountains. ERA5 provides higher P amounts over all five macroclimatic zones, while RF-MEP generally yields the lowest annual P values. Over the Far North, all products show a marked rainy season during December–March due to summer convective P, which differs from the marked seasonality evident over the Near North, Central Chile, and South regions. Over the Far North, ERA5 presents the highest mean annual P (157 mm), which is almost twice the amount provided by the second-highest product MSWEPv2.8 (83 mm), this is followed by CR2MET (63 mm), while RF-MEP has the lowest mean annual P (40 mm). Although ERA5 presents the highest mean annual P values over the Near North, Central Chile, and South regions (208 mm, 902 mm, and 2172 mm, respectively), when considering only our case study catchments (Figure 2b), CR2MET has the highest mean monthly values over the Central Chile and South regions during April–June. RF-MEP and MSWEPv2.8 have similar mean annual P values over Central Chile (670 mm for both products) and the South (1670 mm and 1735 mm, respectively) regions, although RF-MEP consistently shows the largest monthly P amounts of the two products over the corresponding catchments. ERA5 provides the highest mean annual P values over the Far South (3,018 mm), followed by CR2MET (1888 mm), MSWEPv2.8 (1714 mm), and RF-MEP (815 mm). Finally, each product shows low seasonality over the Far South. Here, ERA5 presents higher monthly P values throughout the year, with the largest difference from the other products between January–March and September–December."

To include a more detailed analysis of the selected P products, we computed the median annual values of four extreme indices (Climdex) for the full time period, as recommended by the Expert Team on Climate Change Detection and Indices

(ETCCDI). Figure 3 of the updated manuscript (Figure R5 below) displays the spatial distribution of the median values, obtained from the entire period, for these four annual indices. Equivalent figures for the separate periods of analysis have been added to the supplement. The following text has been added to the manuscript between L172–191 to introduce and describe Figure 3:

"To gain a deeper understanding of the differences between the four P products, we examined the spatial distribution of median annual values of four Climdex Indices (Karl et al., 1999) for 1990–2018 (Figure 3). First, to account for days without rain ($P < 1$ mm), we used the consecutive dry days index (CDD; Figure 3a), which retrieves the maximum dry spell length. It is evident that CR2MET yields longer dry spells, mainly across the Far North and Near North regions, while ERA5 has shorter dry spells over these regions, especially over the Andes Mountains. CR2MET, RF-MEP, and MSWEPv2.8 have similar spatial patterns over the Central Chile and South regions, while ERA5 has less consecutive dry days over the Andes Mountains. Similarly, ERA5 provides shorter dry spells over the Far South, while CR2MET and RF-MEP present similar patterns. These results are consistent with the consecutive wet days index (CWD; Figure 3b), which assesses the frequency and intermittency of P. ERA5 provides the highest CWD values over the driest regions (Far North and Near North), with medians ranging from 0 to 25 days, followed by MSWEPv2.8 (0 to 15 days). ERA5 also shows higher CWD values over high-elevation areas in Central Chile, while the remaining products show similar spatial patterns to each other. The four products show agreement in the CWD over the South region, with values ranging from 5 to 25 days. Finally, RF-MEP shows the lowest consecutive days with P in the Far South, followed by CR2MET and MSWEPv2.8, while ERA5 shows substantially higher CWD values at latitudes greater than 47° S.

To characterise high P intensities, we used the Rx5day (Figure 3c) and R95pTOT (Figure 3d) indices, which represent the maximum P accumulated over five consecutive days, and the total P above their 95th percentile of the daily P for wet days, respectively. Figure 3c shows that ERA5 and CR2MET generally yield the highest Rx5day values, followed by MSWEPv2.8 and RF-MEP. A similar spatial variability is obtained with R95pTOT (Figure 3d), indicating that there is a greater contribution of P from extreme events in ERA5 over high-elevation areas. These spatial patterns are replicated to some extent by CR2MET, which provides R95pTOT values up to 1200 mm over the Andes Mountains in Central Chile."

Similarly, we have replicated this figure for the calibration, Verification 1, and Verification 2 periods (Figure R6, Figure R7 and Figure R8 of this response, respectively), which have been added to the supplement.

Finally, the improved comparison between P products helped us better understand the causes of differences in the modelling results. We have added a Section titled *"How does the calibration of TUWmodel compensate for differences in P?"* (Section 5.2, L495–529), where we discuss how the calibration of TUWmodel compensate for differences between the selected P products.

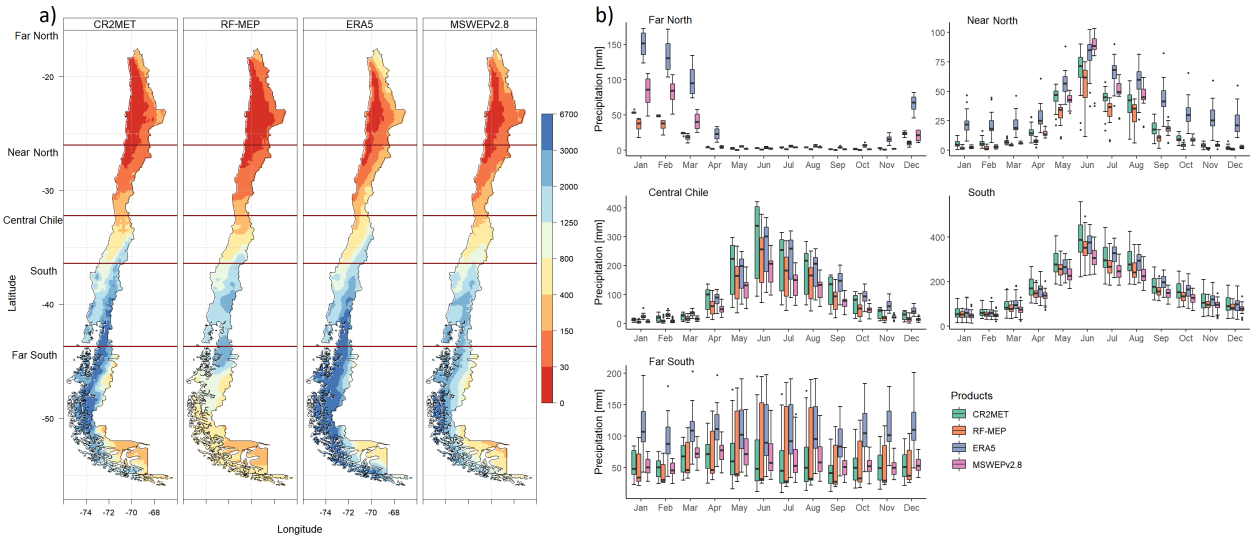

Figure R1. Comparison of P products over 1990–2018 (full time period): a) mean annual P for each product resampled to a 0.05° spatial resolution using the nearest neighbour method. The dark red horizontal lines represent the limits of each major macroclimatic zone; and b) mean monthly P averaged over each catchment located within each macroclimatic zone (see Figure 1d).

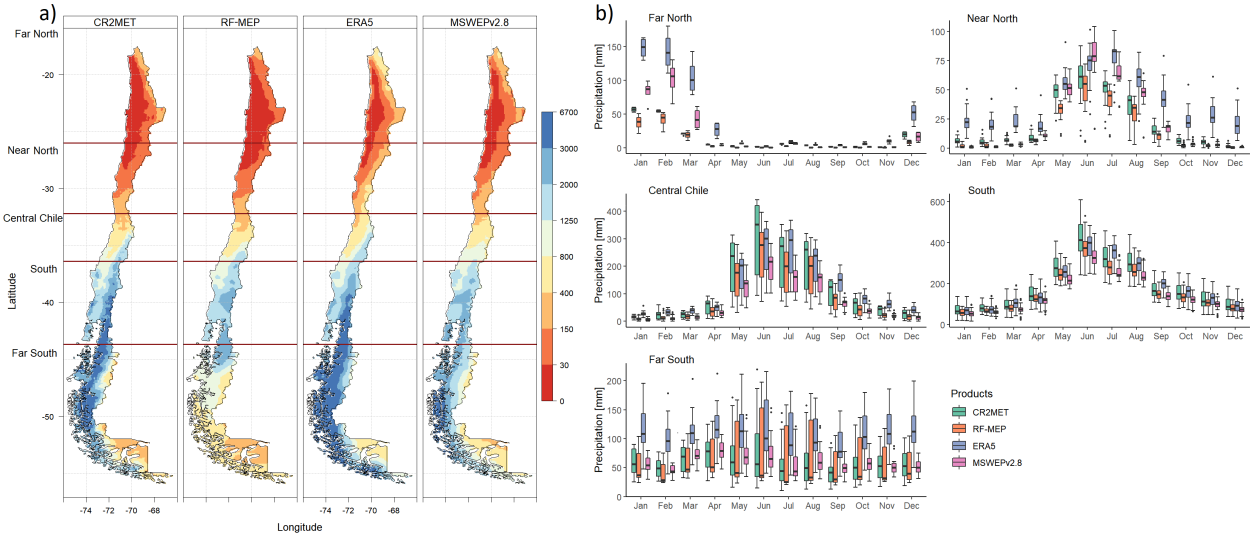

Figure R2. Comparison of P products over 2000–2014 (near-normal): a) mean annual P for each product resampled to a 0.05° spatial resolution using the nearest neighbour method. The dark red horizontal lines represent the limits of each major macroclimatic zone; and b) mean monthly P averaged over each catchment located within each macroclimatic zone (see Figure 1d).

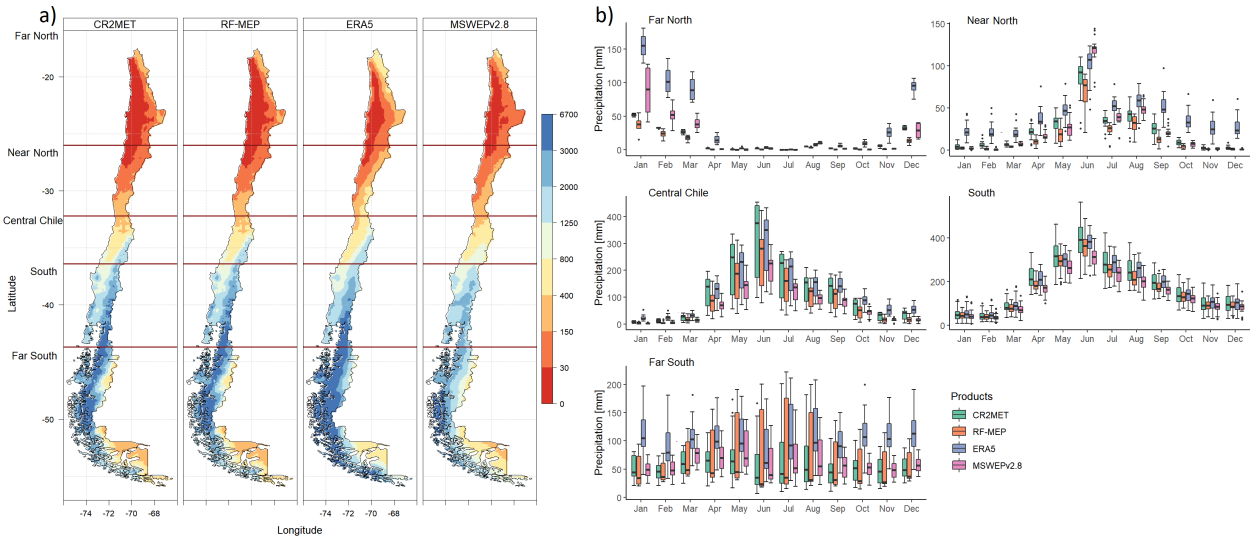

Figure R3. Comparison of P products over 1990–1999 (near-normal): a) mean annual P for each product resampled to a 0.05° spatial resolution using the nearest neighbour method. The dark red horizontal lines represent the limits of each major macroclimatic zone; and b) mean monthly P averaged over each catchment located within each macroclimatic zone (see Figure 1d).

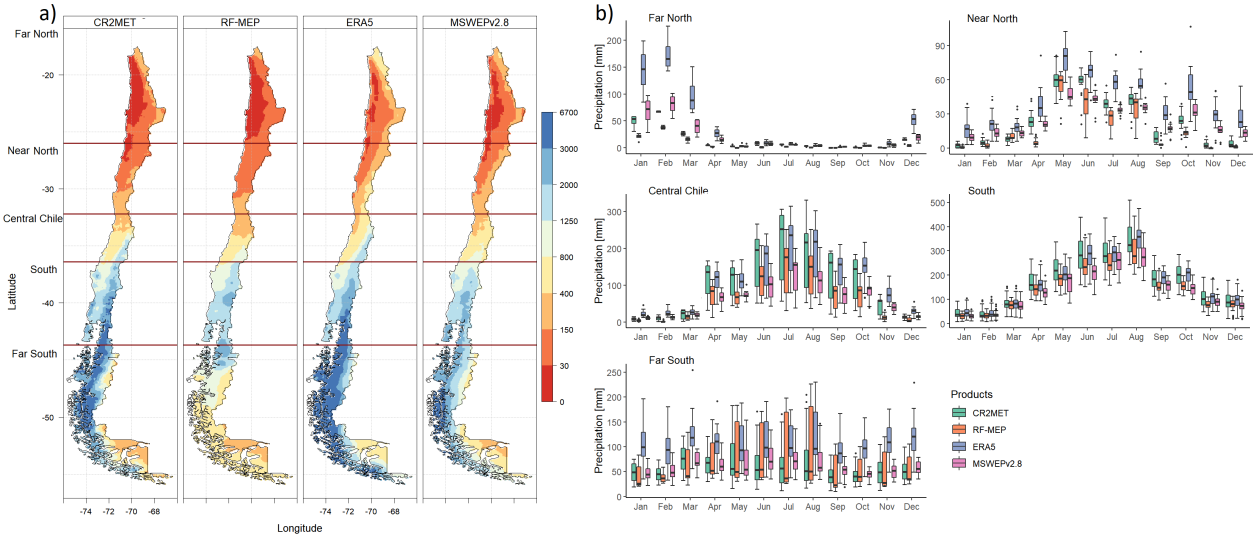

Figure R4. Comparison of P products over 2015–2018 (dry): a) mean annual P for each product resampled to a 0.05° spatial resolution using the nearest neighbour method. The dark red horizontal lines represent the limits of each major macroclimatic zone; and b) mean monthly P averaged over each catchment located within each macroclimatic zone (see Figure 1d).

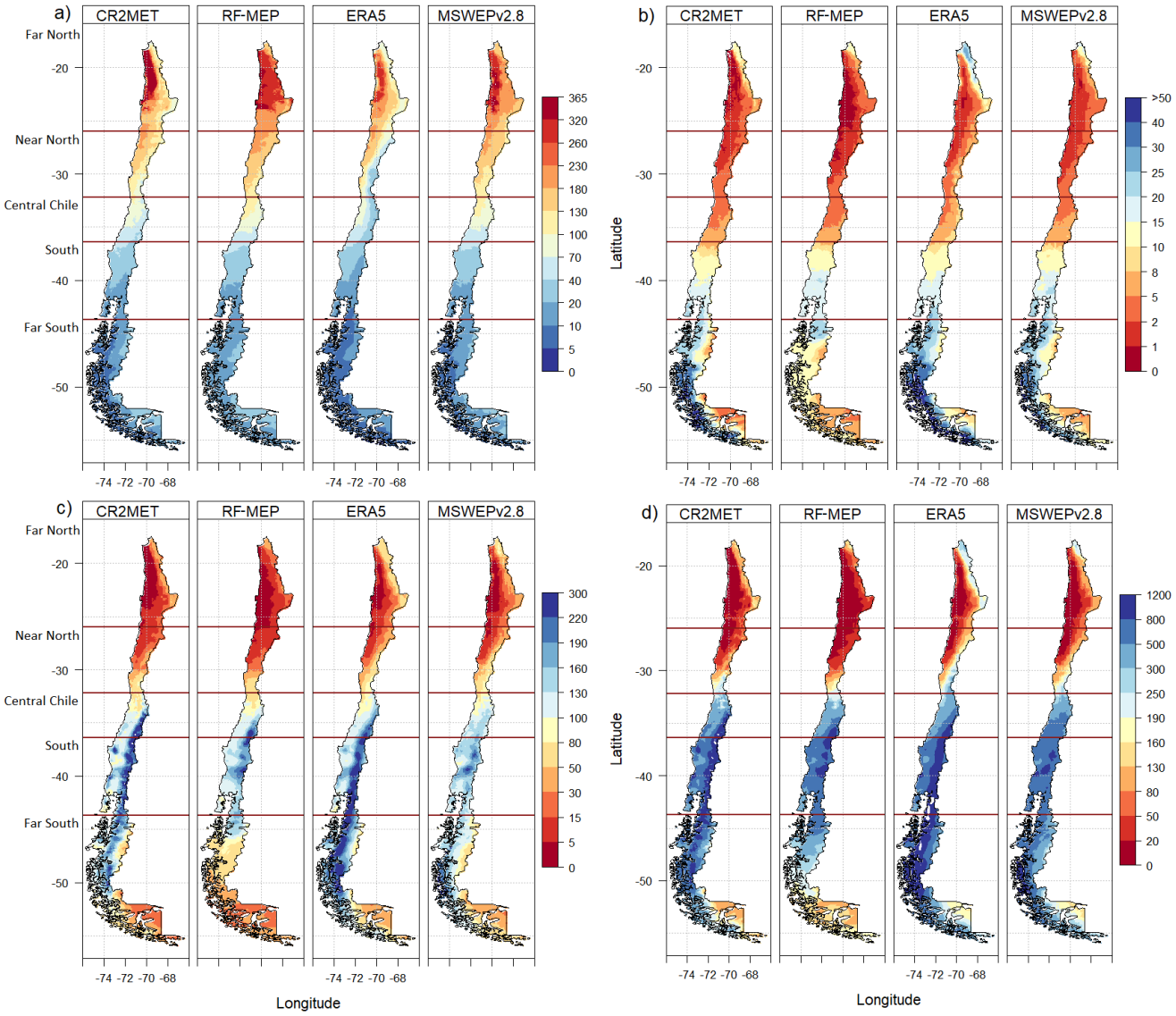

Figure R5. Median annual values of four Climdex indices over 1990–2018 (full period): a) number of consecutive dry days (CDD); b) number of consecutive wet days (CWD); c) maximum P over five consecutive days (RX5day); and d) annual P accumulated for events that are above the 95th percentile of the daily P for wet days (R95pTOT). The dark red horizontal lines represent the limits of each macroclimatic zone.

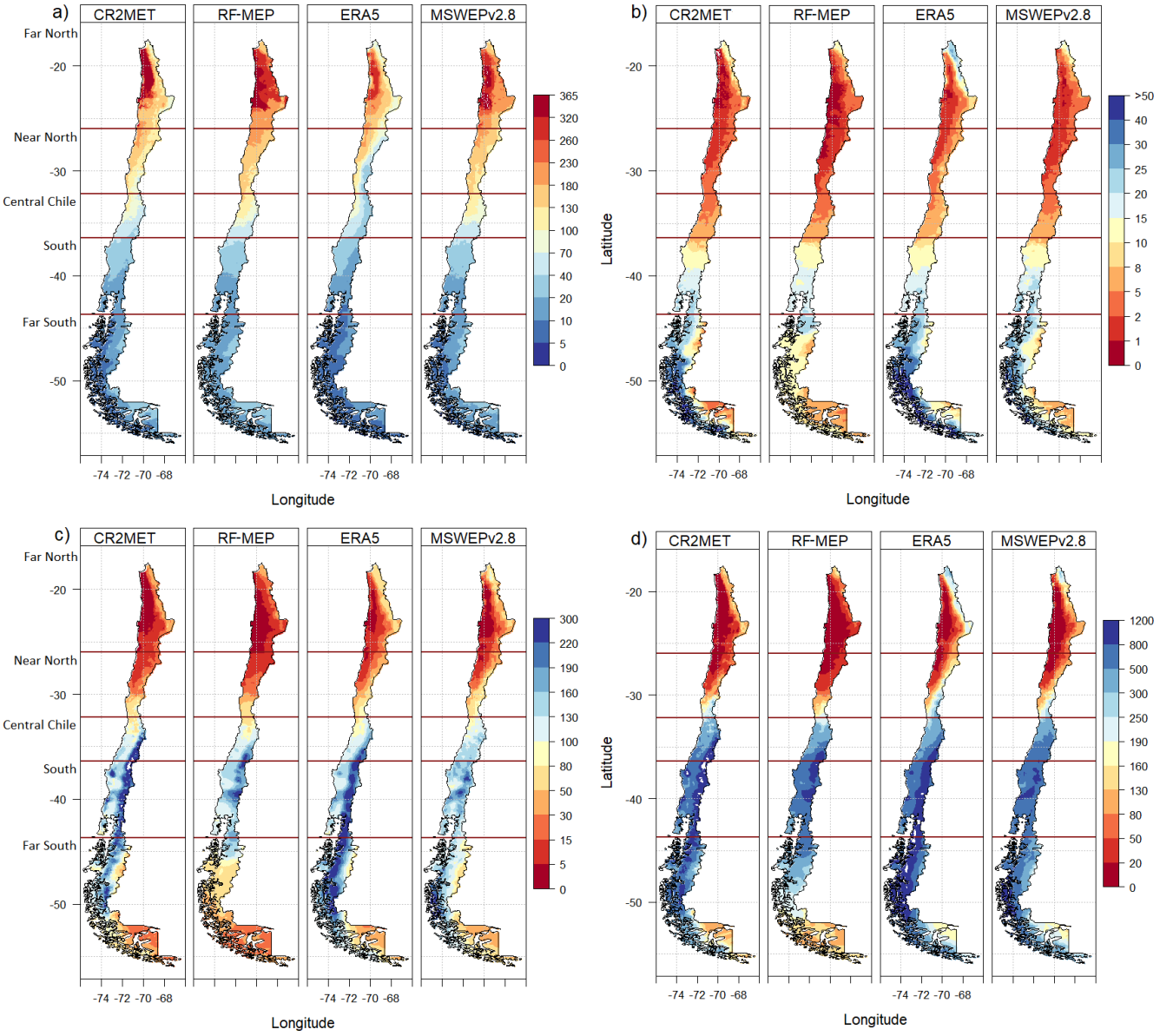

Figure R6. Median annual values of four Climdex indices over 2000–2014 (near-normal): a) number of consecutive dry days (CDD); b) number of consecutive wet days (CWD); c) maximum P over five consecutive days (RX5day); and d) annual P accumulated for events that are above the 95th percentile of the daily P for wet days (R95pTOT). The dark red horizontal lines represent the limits of each macroclimatic zone.

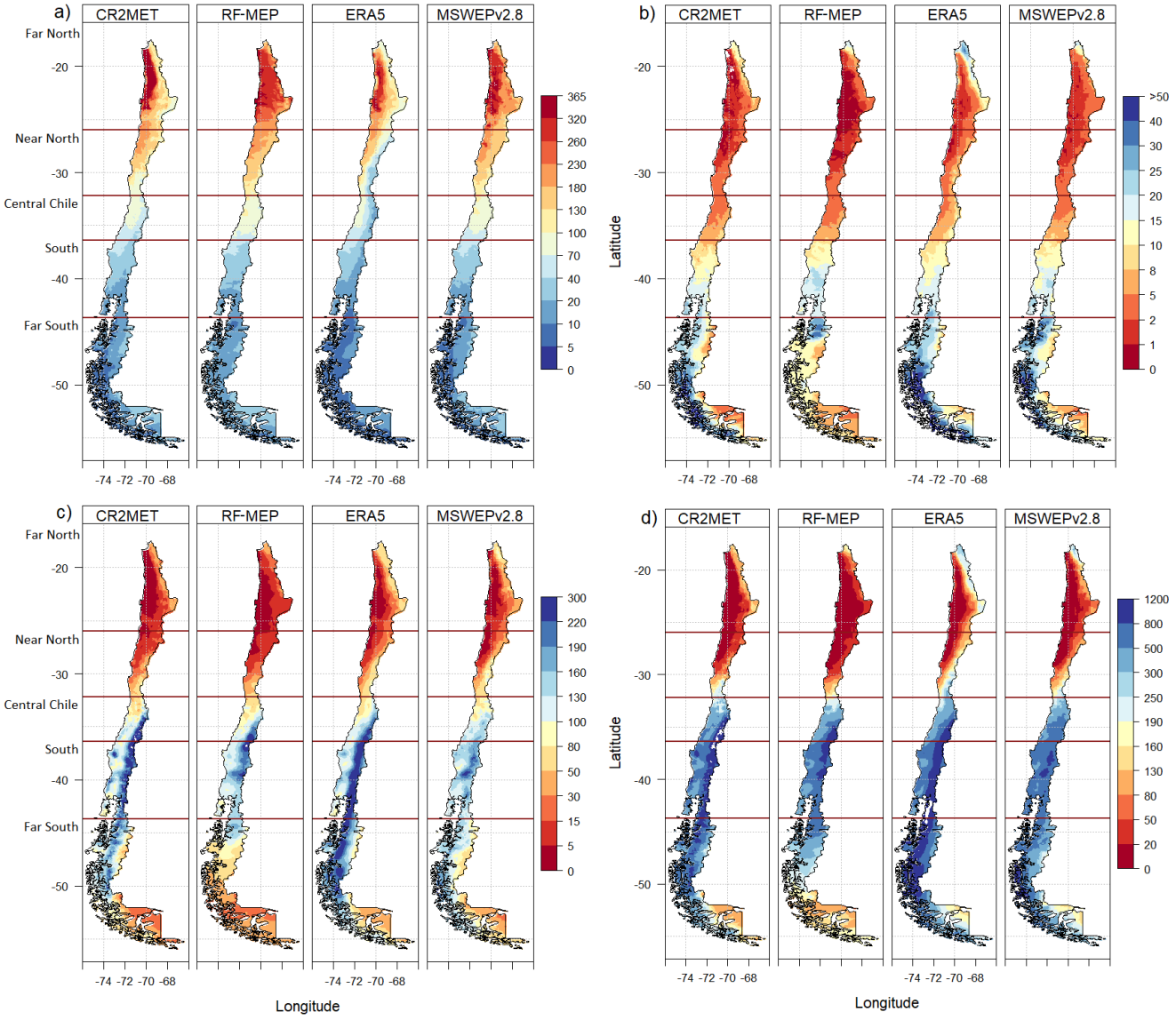

Figure R7. Median annual values of four Climdex indices over 1990–1999 (near-normal): a) number of consecutive dry days (CDD); b) number of consecutive wet days (CWD); c) maximum P over five consecutive days (RX5day); and d) annual P accumulated for events that are above the 95th percentile of the daily P for wet days (R95pTOT). The dark red horizontal lines represent the limits of each macroclimatic zone.

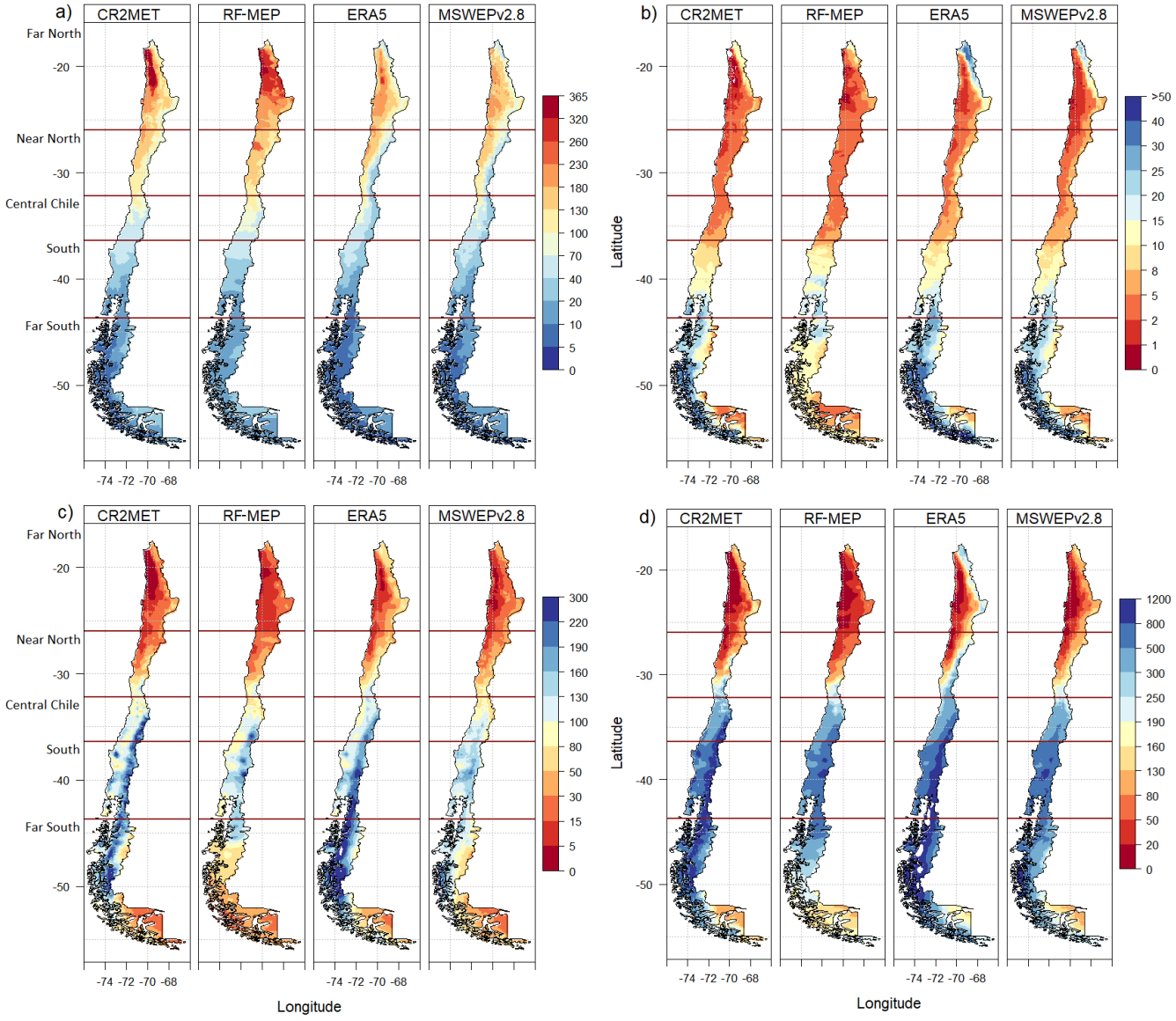

Figure R8. Median annual values of four Climdex indices over 2015–2018 (dry): a) number of consecutive dry days (CDD); b) number of consecutive wet days (CWD); c) maximum P over five consecutive days (RX5day); and d) annual P accumulated for events that are above the 95th percentile of the daily P for wet days (R95pTOT). The dark red horizontal lines represent the limits of each macroclimatic zone.

JPC2: The testing of regionalisation performance for the full period of observations (1990-2018) somewhat mixes the interpretations. I like the idea of splitting the period into calibration and two validation periods showing different climate conditions. So, testing the regionalisation model performance for the same periods can bring some additional information about the performance of different methods in different climate conditions and also show the loss in regionalisation compared to calibration and validation efficiencies.

We thank Juraj Parajka for raising a very relevant point. For brevity, we presented the regionalisation results only for the entire period of 1990–2018 in the first version of the manuscript. However, we agree that evaluating the performance of regionalisation techniques for the same periods used for calibration and verification of the individual catchment performances bring additional information by providing results that are directly comparable to those presented in Section 4.1.1 (Calibration and verification), and help us to understand the regionalisation performance for near-normal and dry conditions.

We evaluated the performance of the three regionalisation methods over the calibration (2000–2014), Verification 1 (1990–1999), and Verification 2 (2015–2018) periods, and modified the manuscript accordingly. To this end, we have included Figure 6 to the manuscript (replicated below as Figure R9), while Figure 7, 8, and 9 also split the performance according to these periods. Furthermore, we have included the following text to the Results Section between L349–365:

"Figure 6 summarises the leave-one-out cross-validation results obtained from the application of three regionalisation methods, for each P product. The results are displayed for the calibration (2000–2014; panel a), Verification 1 (1990–1999; panel b), and Verification 2 (2015–2018; panel c) periods. Overall, the median performance of all P products was the best for feature similarity, with median KGE' values between 0.44–0.62 for all periods, followed by spatial proximity (0.39–0.55) and parameter regression (-0.12–0.51). In addition to exhibiting a considerably lower overall performance, parameter regression returned a larger spread in KGE's for all periods.

The overall performances obtained for feature similarity and spatial proximity are relatively close for different P products over each period (Figure 6). For feature similarity, all P products generate acceptable KGE' results (median KGE' $>$ 0.54) during the calibration and Verification 1 periods, while the median KGE' values during the dry Verification 2 period lowered to a median KGE' of $>$ 0.44. The best model performance for feature similarity was obtained by CR2MET, with median KGE' values of 0.62 for calibration and Verification 1, and 0.53 for Verification 2, followed closely by RF-MEP for calibration (0.59), ERA5 for Verification 1 (0.59), and MSWEPv2.8 for Verification 2 (0.52). In the case of spatial proximity, MSWEPv2.8 yielded the best performance in the calibration period (0.55), followed closely by RF-MEP (0.56, but with a higher dispersion), and CR2MET (0.53). For Verification 1, RF-MEP provided the best performance (0.54), while MSWEPv2.8 produced the best results over Verification 2 (0.48). For spatial proximity, ERA5 performed the worst over the three evaluated periods. Finally, parameter regression yielded the lowest results, with CR2MET and ERA5 showing the highest median KGE' values ($>$ 0.42 for calibration and Verification 1, and $>$ 0.22 for Verification 2)."

Furthermore, we have adapted the Discussion Section to discuss regionalisation performances over the three period. This is most evident in Section 5.1 (*"Performance of P products"*) and Section 5.3 (*"Evaluation of regionalisation techniques"*).

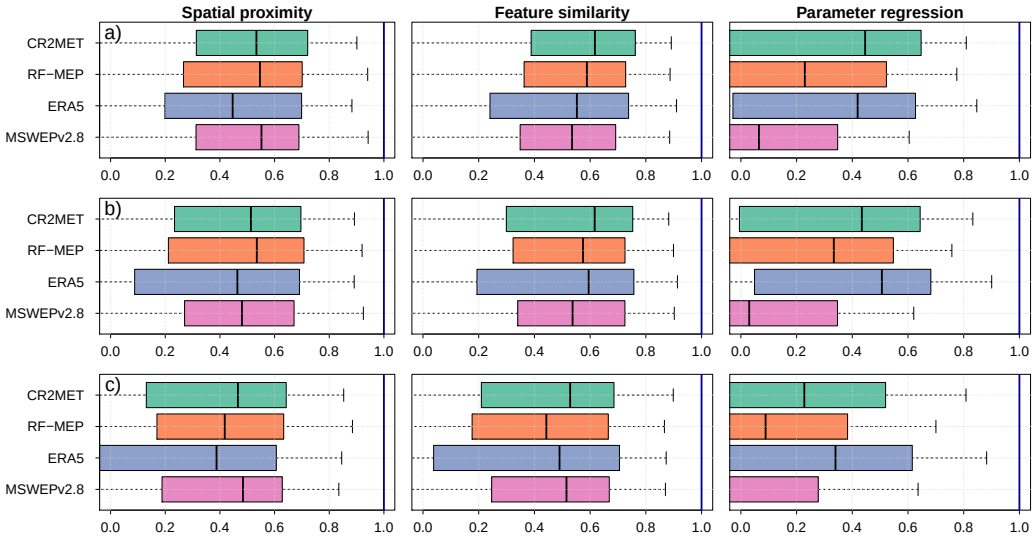

Figure R9. Leave-one-out cross-validation results for the three regionalisation methods applied with different P products during the: a) calibration (2000–2014); b) Verification 1 (1990–1999); and c) Verification 2 (2015–2018) periods.

JPC3: The results indicate that the calibration procedure can compensate for some differences in precipitation products. It will be interesting to see which model parameters/runoff generation processes are affected and whether and how the parameter values change between different hydrological regimes. Are the other simulated components of water balance similar as well (between the products)?

We thank Juraj Parajka and Reviewer 2 (R2C1) for bringing up this important point. We agree that it is important to demonstrate how the calibrated model parameters compensate for differences between P products. Figures R10, R11, R12 and R13 show the distribution of model parameters obtained through model calibration according to hydrological regime, while Figures R14, R15, R16, and R17 show the respective monthly water balance components. Figures R11 and R15 (nivo-pluvial regime) have been added to the manuscript (Figures 12 and 13), while the remaining figures have been added to the supplement. We added the following text between L496–522 to a new section titled *"How does the calibration of TUWmodel compensate for differences in P?"*:

"The calibration of TUWmodel was able to compensate, to some extent, for differences in annual and intra-annual P amounts, intermittency, and extremes (see Figures 2 and 3) among the four products. Using the example of the nivo-pluvial catchments, Figure 12 illustrates how TUWModel parameters compensate for differences between the P forcings used in

calibration, while Figure 13 shows the corresponding variations in the mean monthly water balance components. Similar figures for snow-dominated, pluvio-nival, and rain-dominated catchments can be found in the supplement (Figures S12–S17).

In general, the calibrated parameters behave as expected for each hydrological regime. A notable exception is ERA5, which shows low values for the snow correction factor (SCF) in nivo-pluvial and snow-dominated catchments (Figures 12 and S12). These catchments are primarily located in the arid Near North region (see Figure **??** *and Figure S15), where the estimated winter P is substantially lower for CR2MET, RF-MEP, and MSWEPv2.8, and a high SCF corrects this apparent underestimation. The lower P amounts presented in these products may reflect the incorporation of information from rain gauges located in drier, low-lying areas to correct their P estimates (see Figure S1).*

ERA5 presented relatively low SCF values over nivo-pluvial catchments compared to the other P products (Figure 12), which is expected because it exhibits the highest P values. Conversely, because RF-MEP has the lowest mean monthly P over the nivo-pluvial catchments, the model adjusts the evaporation, snow water equivalent, and soil moisture components (Figure 13), thus increasing the simulated Q (to match the observed Q). Substantial differences were obtained for LPrat and field capacity (FC), which directly affect evaporation and soil moisture. For example, over the nivo-pluvial catchments, the LPrat and FC values for RF-MEP are similar to those of ERA5, despite RF-MEP having substantially lower P amounts, which in turn is reflected in the reduced soil moisture and evaporation amounts. The differences between LPrat and FC according to P product are even more pronounced for snow-dominated catchments (Figure S12)."

Finally, higher values of the nonlinear parameter for runoff production Beta (Széles et al., 2020, their Eq. 7) reduce the amount of water that leaves the catchment as runoff. For all hydrological regimes except pluvio-nival, the median Beta parameter is substantially higher for ERA5 than for the other P products. The larger Beta values obtained with ERA5 are expected to attenuate the runoff generation from extreme P events (see Figure R5c–d). Interestingly, the Beta parameter is zero in some pluvio-nival catchments, which means that all liquid P and snowmelt was used to generate runoff (Figure S16). This behaviour was more pronounced with RF-MEP and MSWEPv2.8, which exhibited the lowest P amounts and longer dry spells (Figure R5a) over these catchments. In general, the storage components obtained from each P product (computed as the sum of the two deepest reservoirs of the model (see Széles et al., 2020, their Figure 3)) are similar for all four P products."

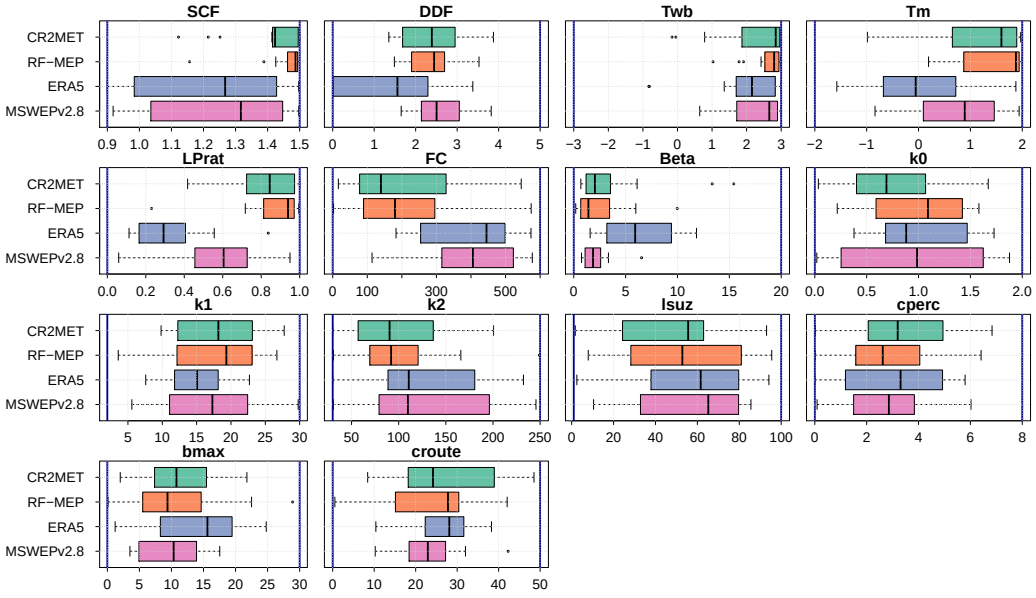

Figure R10. Model parameters obtained through calibration in snow-dominated catchments. The vertical blue lines indicate the upper and lower limits of the parameter ranges.

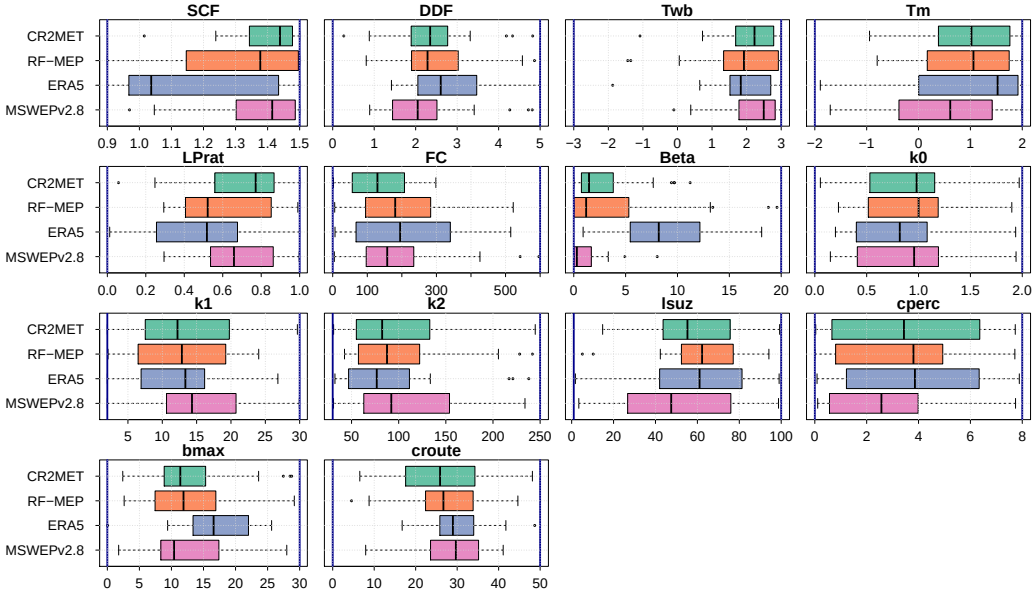

Figure R11. Model parameters obtained through calibration in nivo-pluvial catchments. The vertical blue lines indicate the upper and lower limits of the parameter ranges.

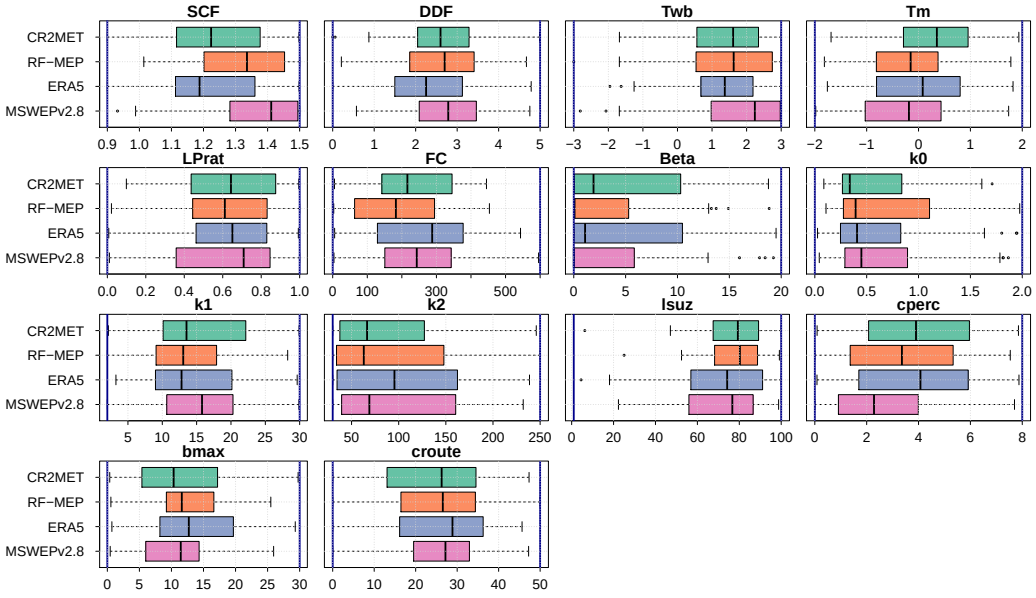

Figure R12. Model parameters obtained through calibration in pluvio-nival catchments. The vertical blue lines indicate the upper and lower limits of the parameter ranges.

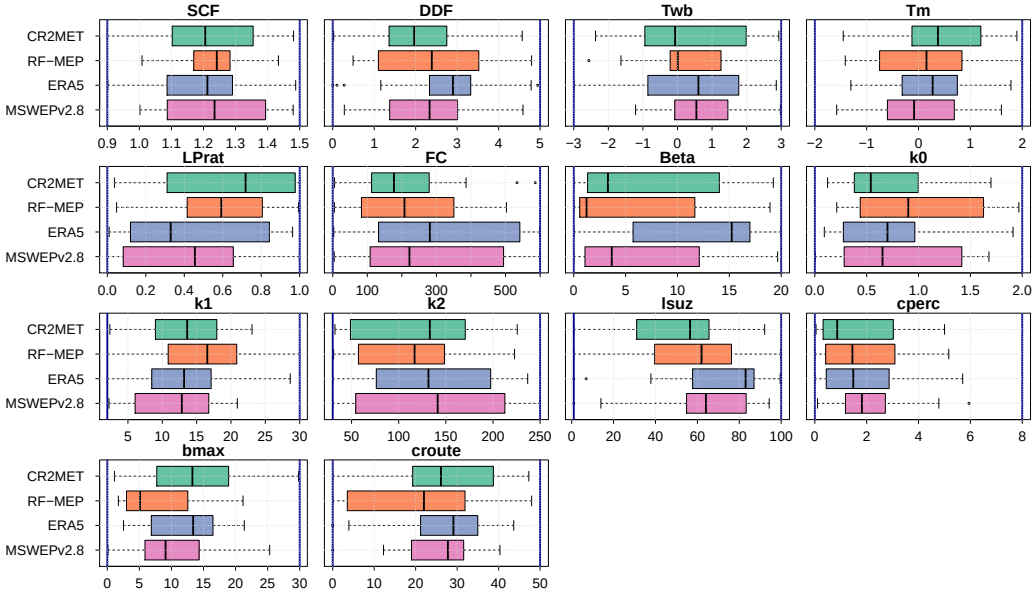

Figure R13. Model parameters obtained through calibration in rain-dominated catchments. The vertical blue lines indicate the upper and lower limits of the parameter ranges.

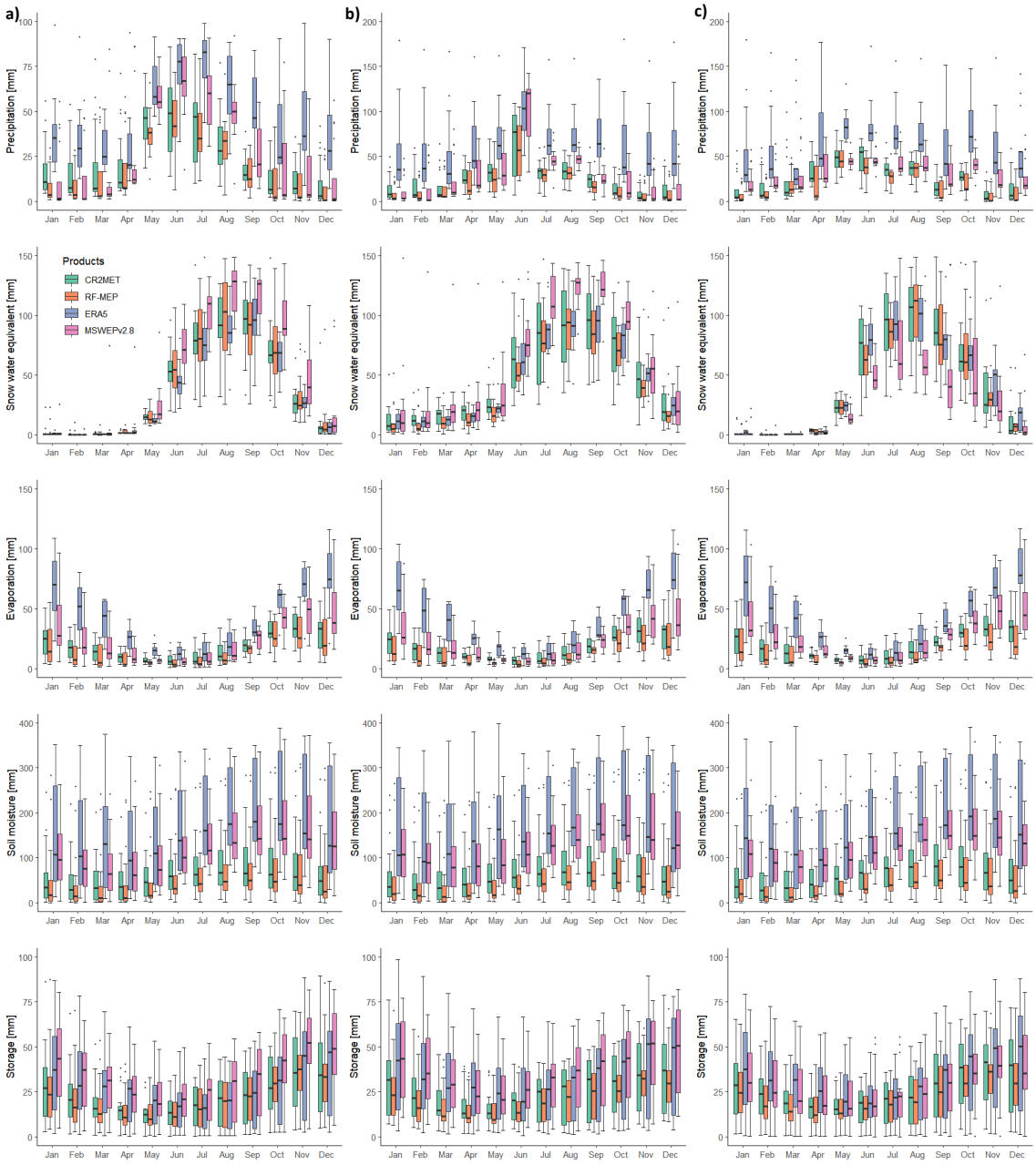

Figure R14. Mean monthly water balance components over snow-dominated catchments, obtained by forcing the TUW model with different P products for the: a) calibration (2000–2014); b) Verification 1 (1990–1999); and c) Verification 2 (2015–2018) periods. The mean monthly P was added for comparison purposes.

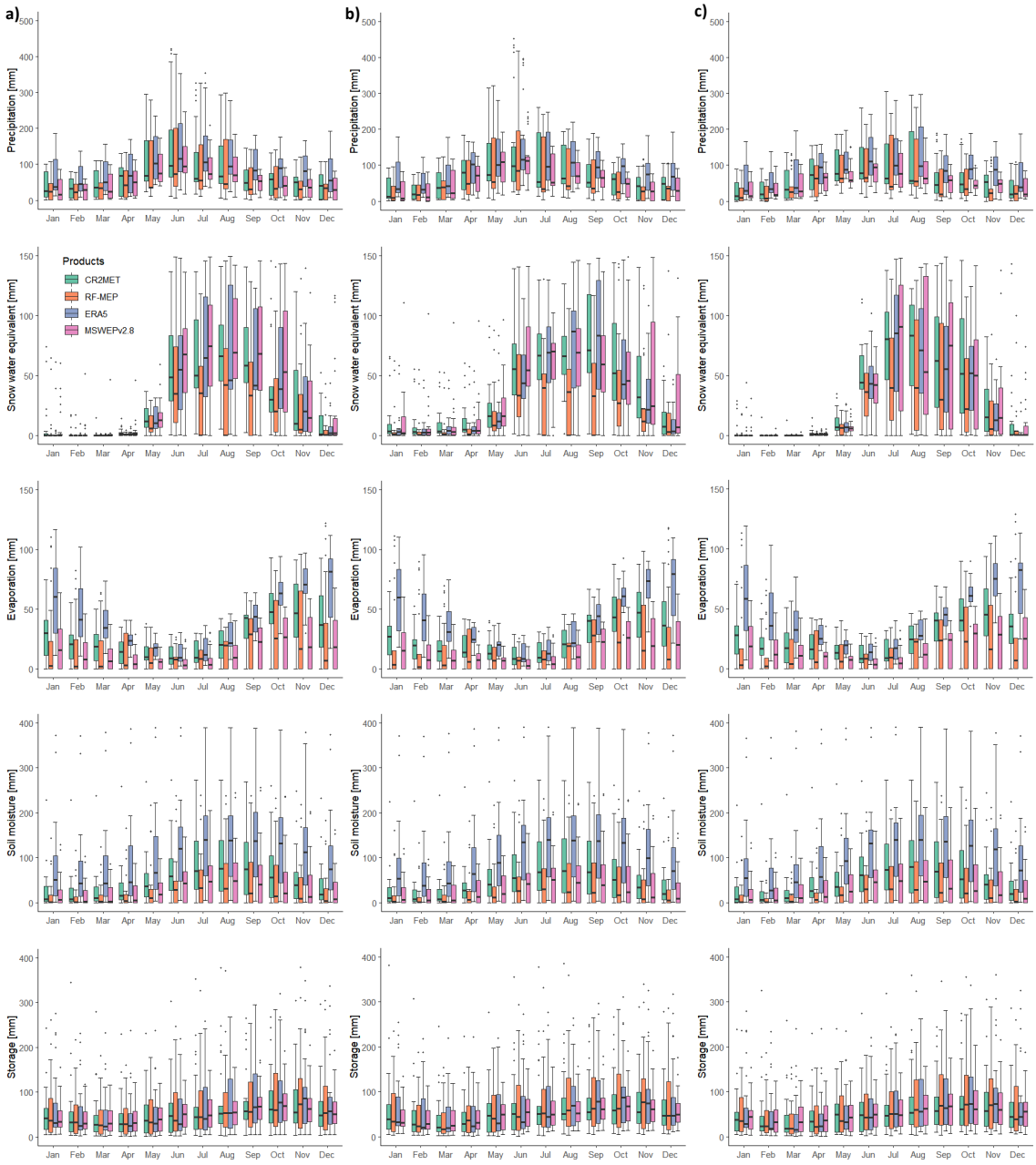

Figure R15. Mean monthly water balance components over nivo-pluvial catchments, obtained by forcing the TUW model with different P products for the: a) calibration (2000–2014); b) Verification 1 (1990–1999); and c) Verification 2 (2015–2018) periods. Mean monthly P was added for comparison purposes.

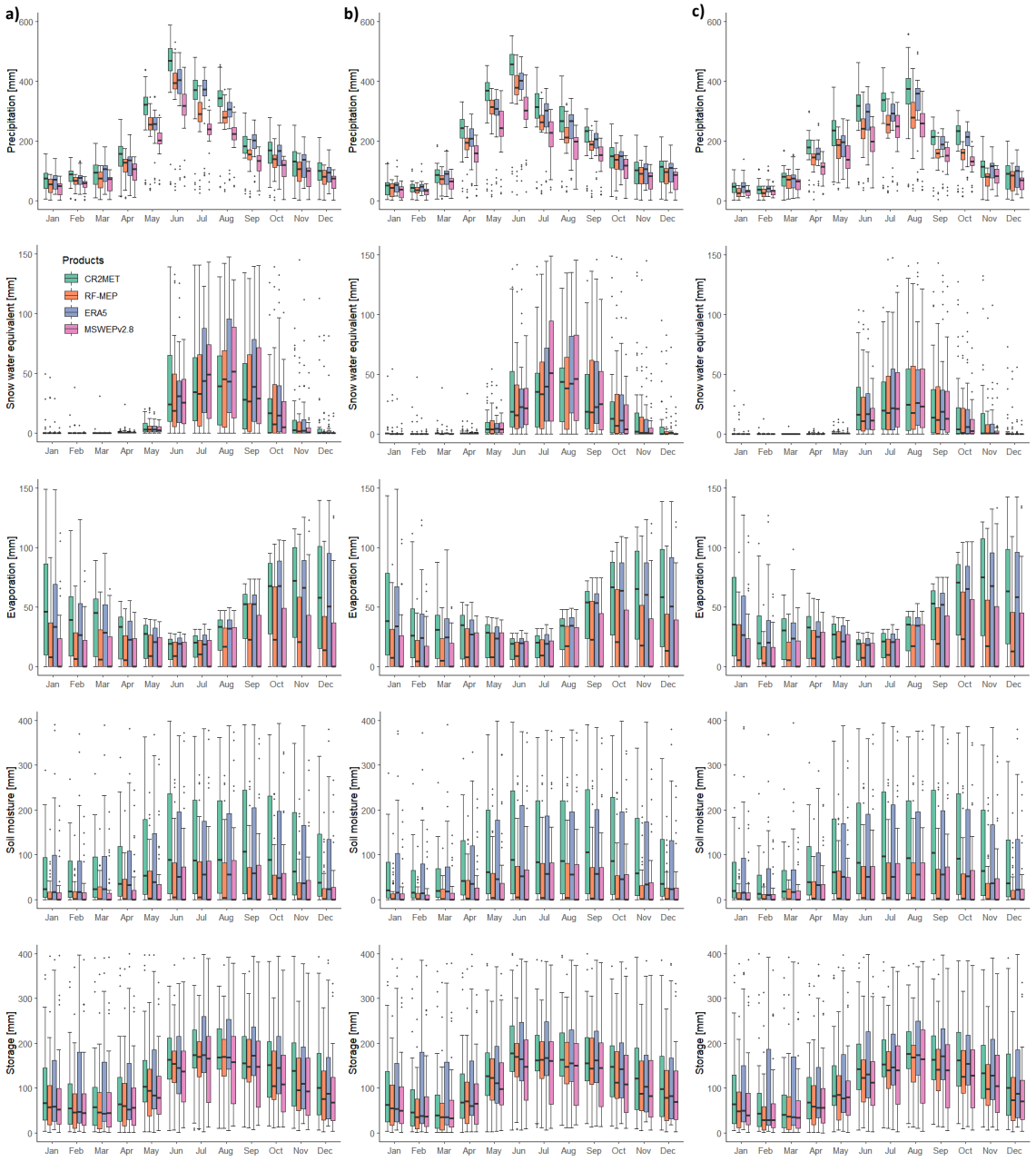

Figure R16. Mean monthly water balance components over pluvio-nival catchments, obtained by forcing the TUW model with different P products for the: a) calibration (2000–2014); b) Verification 1 (1990–1999); and c) Verification 2 (2015–2018) periods. The mean monthly P was added for comparison purposes.

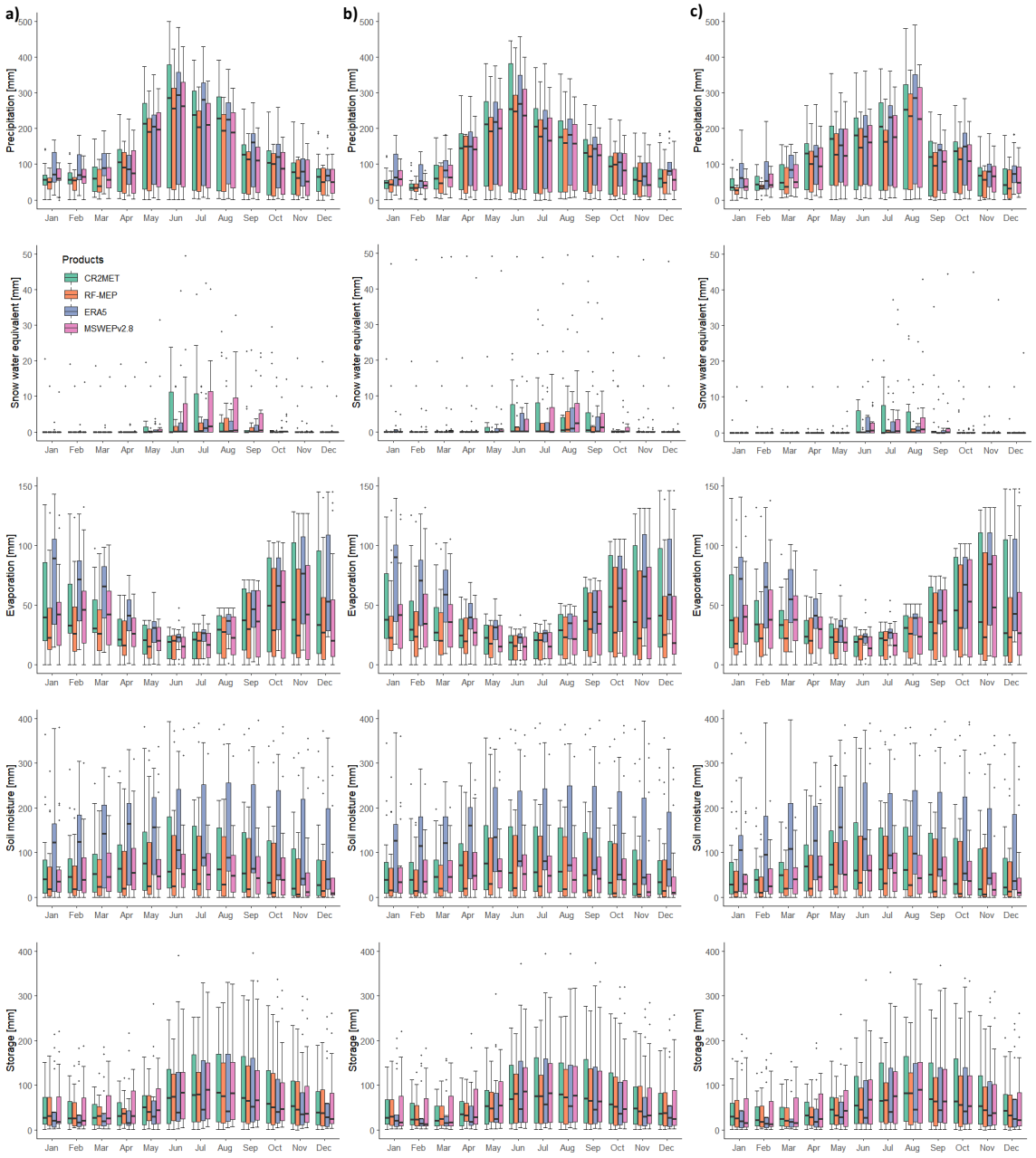

Figure R17. Mean monthly water balance components over rain-dominated catchments, obtained by forcing the TUW model with different P products for the: a) calibration (2000–2014); b) Verification 1 (1990–1999); and c) Verification 2 (2015–2018) periods. The mean monthly P was added for comparison purposes.

JPC4: It will be interesting to see some disentangling difference between similarity and spatial proximity methods for linking the results with previous studies. Is it the averaging of ten simulations or the similarity between the catchment attributes that impact the regionalisation performance? What is the impact of averaging ten simulations compared to simulations based on the most similar catchment only?

Thank you for this comment! We agree that it is interesting to assess the influence of the number of donor catchments for the case of feature similarity. We have added Figure 11 to the revised manuscript (replicated below as Figure R18) and added the following paragraph to the Results Section between L435–443:

"Figure 11 shows the performance of feature similarity during the calibration and both verification periods when varying the number of donors used to transfer model parameters to ungauged catchments (see Section 3.6). In general, the highest median performance is obtained when using 4 or more donor catchments. However, the application of a t-test demonstrated that the improvement in the KGE' values obtained when increasing to more than one donor was not statistically significant. The results show that the performance varies according to the P product and selected period of analysis. For the calibration period, feature similarity produced similar median values to those obtained with spatial proximity when one donor was used, while the performance improved as more donors were included. For both verification periods, feature similarity (median KGE' values from 0.44 to 0.64) outperformed spatial proximity (median KGE' values ranging 0.39 to 0.54). For all three periods, feature similarity provided better performance considering the distribution of the KGE' values."

Additionally, we added the following paragraph to the Discussion Section between L571–579:

"Increasing the number of donor catchments in feature similarity improved the regionalisation performance. This is in agreement with several studies that have demonstrated that using an ensemble of multiple donor catchments improves regionalisation results (McIntyre et al., 2005; Zelelew and Alfredsen, 2014; Garambois et al., 2015; Beck et al., 2016; Neri et al., 2020). Figure R18 shows that there is a slight increase in performance when 4 donors or more are used, independent of the P product and evaluated period. These results are similar to those of Neri et al. (2020), who determined that three donors were optimal for the TUWmodel over Austrian catchments. Feature similarity still outperformed spatial proximity when only one catchment was used to transfer the model parameters to the ungauged catchments, which is in agreement with multiple studies that have shown the ability of this method to produce good regionalisation results (Parajka et al., 2005; Oudin et al., 2008; Bao et al., 2012; Garambois et al., 2015; Neri et al., 2020)."

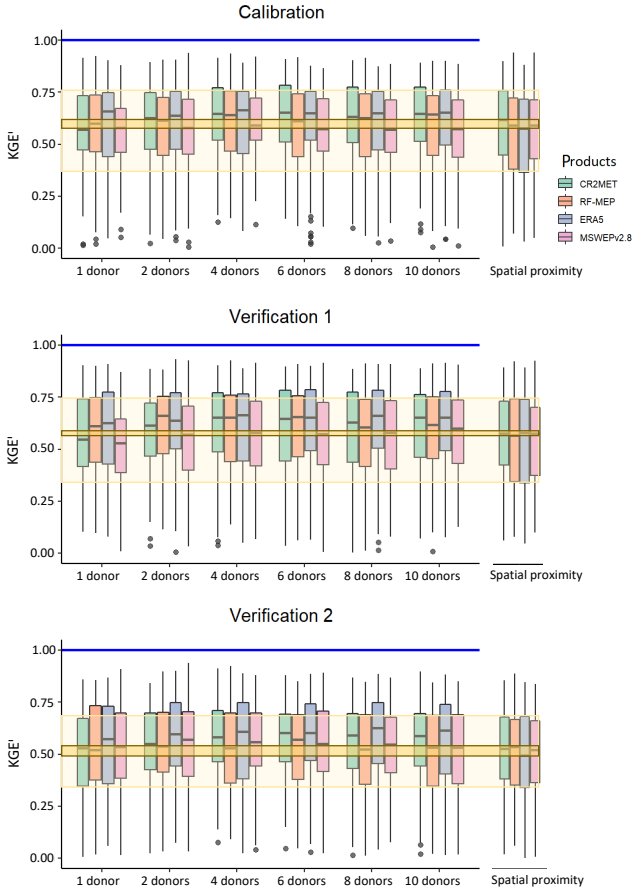

Figure R18. Influence of the number of donors used for feature similarity for calibration (2000—2014); Verification 1 (1990–1999); and Verification 2 (2015–2018). The results from spatial proximity are included on the right of each panel for comparison purposes. The dark yellow box denotes the upper and lower bounds of the median performance (of the four P products) obtained with spatial proximity, while the lighter yellow box represents the upper an lower bounds of the interquartile range for spatial proximity.

JPC5: Table 1: Leave-one-out and jackknife cross-validation are likely the same procedure. Please consider using, for consistency, only one term for the same procedure.

Thank you very much for this observation. We have decided to use the term "leave-one-out" consistently throughout the manuscript as suggested, thus avoiding any potential confusion.

Reviewer 2 (R2)

R2–General comments: The study investigates the impact of four different gridded precipitation products (ERA5, MSWEPv2.8, RF-MEPv2, and CR2MET) having different spatial resolutions on the relative performance of three regionalization techniques (spatial proximity, feature similarity and model parameter regression) over 100 near-natural catchments in Chile, characterized by varying topography, climate and hydrologic regimes. TUWmodel, a semi-distributed HBV-like model, was driven separately by four precipitation products and calibrated/evaluated for each catchment using two performance metrics (KGE' and AOF) utilizing daily streamflow. The authors compared the performance of the regionalisation techniques through a leave-one-out cross-validation exercise, which consists of leaving out each one of the 100 catchments, transferring model parameters, conducting flow simulations and computing performance metrics. They concluded that the calibration procedure is able to compensate for the differences in precipitation products and the spatial resolution of the precipitation product does not largely affect the regionalization performance. Overall, feature similarity provided the best regionalization performance followed closely by spatial proximity, while parameter regression performed the worst. They also reported that the hydrologic regime impacts the performance of regionalization.

I would like to thank the authors for conducting this interesting work. The topic of the manuscript fits well to the journal scope and readership. The use of language and structure is well. The authors put significant effort to summarize the outcome of substantial number of cases (four precipitation products, three regionalization techniques, 2 objective functions as well as hydrological regimes) into a single manuscript. However, it is still difficult for the reader to follow through and understand the full reasoning behind some of the study outcomes. My suggestion for improving the manuscript would be to explain the link between different precipitation products and calibration/verification and regionalization methods through differences they make in model components (parameters/fluxes) and components of the efficiency metrics (KGE and flow duration curve segments). I listed my main and minor suggestions for improvement/revision in the sections below.

We thank the reviewer for their constructive comments. We have addressed all points raised and believe that the manuscript has benefited substantially from these comments. Please see below for the detailed responses to their comments.

R2C1: Discussion of how the model components compensated for differences in precipitation products: There are significant differences in precipitation products (for example ERA5 reported four times higher precipitation in the dry north), however most of the calibrated model parameters have similar distribution for ERA5 and other products (Figures S4 and S5), except SCF, Lprat, FC, Beta (I am surprised to see that parameter distributions barely touch the limits). An analysis of how the model compensated for differences in precipitation products through parameter values and fluxes other than streamflow will significantly improve the reasoning behind study outcomes. For example, did the large bias in ERA5 in dry regions compensated by more evapotranspiration/groundwater loss etc? The authors also mentioned this topic in Lines 64-67: "Although hydrological model calibration can partly compensate for errors in the representation of P, this may lead to unrealistic model

behavior, thus affecting the quality of parameter regionalisation results" and in Lines 457-459: "The equifinality of model parameters may also impact the relative performance of the regionalisation techniques by producing unrealisitic parameter sets, particularly for the case of parameter regression." Therefore, providing answers to these statements will significantly improve the manuscript.

Reviewer 2 raises some very important points that were not sufficiently covered in the previous version of the manuscript. Firstly, in the revised manuscript, we compare the seasonality, dry and wet spells, and extreme values of the four P products over the five major macroclimatic zones of Chile (please see our response to JPC1).

We agree that it is important to understand how the model compensated for differences between P products. To accomplish this, we evaluated the distribution of model parameters and water balance components per hydrological regime (as opposed to boxplot showing the distributions for all 100 catchments in the initial version of the manuscript). Please refer to our response to JPC3, where we have addressed how the model was able to compensate for differences in P forcings, including discussing the specific case of ERA5 over dry regions. The fact that parameter distributions barely touch their limits very likely is due to the calibration algorithm used for this study (SPSO-2011), although a comparison with other calibration algorithms is beyond the scope of this article.

Finally, now that we have more comprehensively analysed how the TUWmodel compensates for differences between P products, the previous text has now been replaced between L539–541 with: *"The compensation for P differences obtained through model calibration also affected the relative performance of regionalisation techniques, producing unrealistic parameter sets in some donor catchments. In particular, such compensation may have impacted the spatial transferability of model parameters with the parameter regression method."*

R2C2: The authors utilized KGE' and signature-based objective functions for calibration, both of which can be decomposed into hydrologically meaningful components to understand differences in model performance driven by different precipitation products as well as in different hydrological regimes. An analysis of how well different hydrograph characteristics (variability, bias, low flows, high flows etc) are reproduced through regionalization will significantly improve the manuscript.

The reviewer raises a very relevant point that can help us gain a greater understanding of differences in model performance. We have added tables to the Appendix that list the quantiles 0.25 and 0.75 of the three components of the KGE' (r, β and γ) for each P product, individual calibration and validation, and each regionalisation method. Note that after the removal of the AOF from the manuscript (for more information, see ET-General comments (Part 3) and ETC5), there is only the KGE' that needs to be discussed.

The following text has been added to the manuscript to refer to these results:

- L447–449: *"When decomposing the results of the KGE' objective function into its three components (see Appendix C), r exhibited the lowest performance, while and γ values were generally closer to their optimal values, particularly for calibration and Verification 1."*

– L455–458: *"... [the] ability of the rainfall-runoff model to compensate for the P forcing (visible in the performances of the β and γ components; Appendix C); and iii) [the] fact that P products still have errors in the detection of P events that could impact the representation of the modelled Q dynamics (as suggested by the lower performance of the r component of the KGE')"*

– L464–466: *"The decomposition of the KGE' into its components also demonstrated the ability of the TUWmodel to compensate for the total volume of P, as the β component was close to the optimum value, particularly for calibration and Verification 1 (see Appendix C)..."*

Table 1. Quantiles 0.25 and 0.75 of the correlation coefficient (r) of the KGE' over the selected catchments.

Pearson's correlation (r)	CR2MET	RF-MEP	ERA5	MSWEPv2.8
Calibration (cal.)	0.78–0.90	0.77–0.88	0.71–0.86	0.77–0.88
Verification 1 (Ver. 1)	0.74–0.88	0.72–0.87	0.67–0.87	0.69–0.86
Verification 2 (Ver. 2)	0.68–0.86	0.59–0.85	0.59–0.86	0.67–0.85
Spatial proximity (cal.)	0.70–0.87	0.68–0.84	0.57–0.82	0.66–0.84
Spatial proximity (Ver. 1)	0.66–0.86	0.63–0.84	0.61–0.84	0.62–0.84
Spatial proximity (Ver. 2)	0.61–0.83	0.51–0.82	0.56–0.83	0.59–0.82
Feature similarity (cal.)	0.74–0.89	0.71–0.88	0.69–0.85	0.72–0.88
Feature similarity (Ver. 1)	0.69–0.88	0.70–0.88	0.67–0.88	0.69–0.86
Feature similarity (Ver. 2)	0.64–0.87	0.59–0.85	0.64–0.87	0.65–0.84
Parameter regression (cal.)	0.54–0.80	0.54–0.69	0.60–0.82	0.42–0.63
Parameter regression (Ver. 1)	0.58–0.80	0.50–0.68	0.64–0.86	0.43–0.62
Parameter regression (Ver. 2)	0.50–0.79	0.43–0.65	0.59–0.84	0.37–0.57

Table 2. Quantiles 0.25 and 0.75 of the bias ratio (β) of the KGE' over the selected catchments.

Bias ratio (β)	CR2MET	RF-MEP	ERA5	MSWEPv2.8
Calibration (cal.)	0.95–0.99	0.93–1.01	0.97–1.02	0.90–1.02
Verification 1 (Ver. 1)	0.89–1.03	0.84–1.02	0.90–1.12	0.77–1.04
Verification 2 (Ver. 2)	0.96–1.19	0.86–1.11	1.00–1.25	0.74–1.06
Spatial proximity (cal.)	0.73–1.09	0.70–1.15	0.74–1.22	0.70–1.13
Spatial proximity (Ver. 1)	0.72–1.12	0.70–1.12	0.72–1.22	0.69–1.08
Spatial proximity (Ver. 2)	0.73–1.30	0.73–1.23	0.77–1.46	0.68–1.14
Feature similarity (cal.)	0.81–1.19	0.78–1.29	0.81–1.35	0.68–1.3
Feature similarity (Ver. 1)	0.80–1.17	0.74–1.24	0.80–1.36	0.69–1.29
Feature similarity (Ver. 2)	0.86–1.40	0.77–1.40	0.86–1.57	0.69–1.27
Parameter regression (cal.)	0.99–2.04	0.89–1.72	0.76–1.78	0.82–3.07
Parameter regression (Ver. 1)	0.99–1.73	0.87–1.65	0.76–1.62	0.83–2.64
Parameter regression (Ver. 2)	1.10–2.05	0.90–1.83	0.88–1.94	0.83–2.54

Table 3. Quantiles 0.25 and 0.75 of the variability ratio (γ) of the KGE' over the selected catchments.

Variability ratio (γ)	CR2MET	RF-MEP	ERA5	MSWEPv2.8
Calibration (cal.)	0.97–1.00	0.95–1.00	0.95–1.01	0.96–1.01
Verification 1 (Ver. 1)	0.93–1.07	0.92–1.06	0.93–1.07	0.93–1.11
Verification 2 (Ver. 2)	0.92–1.13	0.91–1.17	0.91–1.12	0.79–1.05
Spatial proximity (cal.)	0.84–1.20	0.84–1.23	0.88–1.24	0.88–1.22
Spatial proximity (Ver. 1)	0.89–1.24	0.84–1.30	0.85–1.32	0.86–1.27
Spatial proximity (Ver. 2)	0.88–1.34	0.85–1.37	0.85–1.38	0.75–1.19
Feature similarity (cal.)	0.74–1.06	0.75–1.06	0.75–1.10	0.78–1.07
Feature similarity (Ver. 1)	0.79–1.04	0.76–1.06	0.77–1.07	0.81–1.03
Feature similarity (Ver. 2)	0.79–1.13	0.75–1.12	0.79–1.15	0.66–0.97
Parameter regression (cal.)	0.80–1.18	1.02–1.50	0.84–1.23	1.26–1.89
Parameter regression (Ver. 1)	0.82–1.20	1.02–1.35	0.87–1.25	1.27–1.69
Parameter regression (Ver. 2)	0.86–1.38	1.15–1.83	0.86–1.46	1.22–1.82

R2C3: Line 226: More information on the optimization algorithm parameters should be provided, such as termination criteria, maximum number of iterations permitted etc. to better understand the calibration procedure and its impact on calibrated parameters.

Thank you for this comment. In the revised version of the manuscript, we have added the following text between L240–242:

"We used the standard PSO 2011 algorithm (Clerc, 2011a, b), defined as spso2011 *in the hydroPSO R package (Zambrano-Bigiarini and Rojas, 2013). We set the number of particles in the swarm (npart = 80), the maximum number of iterations (maxit = 100), and the relative convergence tolerance (reltol =* $1E-10$*)."*

Additionally, readers are referred to a manual on how to use hydroPSO to calibrate the TUWmodel between L245–246:

"For more details on the use of the hydroPSO package to calibrate the TUWmodel, readers are referred to Zambrano-Bigiarini and Baez-Villanueva (2020)."

R2C4: Lines 408-410: ERA5 re-analysis product is also a merged product and uses ground-based measurements (after 2009). Please discuss and revise where necessary. (see Hersbach et al., 2020;https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3803). Possibly focus on differences in gauge density used in correction of different precipitation products.

We thank Reviewer 2 for this comment. We understand merging as a process that involves the direct combination of P from different sources, such as ground-based data and gridded products. As per Beck et al. (2019), we do not consider that ERA5 is a merged product because reanalysis products assimilate different types of meteorological data to produce hydrometeorological fields. Although it is true that ERA5 uses information related to rain rate from radar-gauge composites from 2009 onwards (as it includes information from the NCEP Stage IV P product), this is only available over the conterminous USA, and it does not incorporate rain gauge data over Chile. This information has now been added to the manuscript in L140–142: *"Although ERA5 also assimilates NCEP Stage IV* P *estimates over the conterminous USA, which combine NEXRAD data with in-situ measurements, it does not incorporate information from any ground-based* P *stations over Chile."*

R2C5: Line 123: replace "perform well over small catchments" with "perform well even over small catchments"

Thank you for the recommendation. We have now written *"perform well even over small catchments"*. Please see L146.

R2C6: Line 124: replace "tend perform" with "tend to perform"

Thank you very much for picking on this typo. We have now written *"tend to perform"*. Please see L146.

R2C7: Line 127: The reader would be interested to see how well CR2MET performs compared to ERA5. A few sentences based on the reference will be helpful.

We agree that such a comparison would be helpful to the reader; however, to the best of our knowledge, such a comparison does not yet exist in scientific literature. That being said, we have included a more comprehensive analysis of the differences between all P products (please see the response to JPC1).

R2C8: Line 128: CR2MET: Boisier et al (2018) states that CR2MET is produced from ERA-Interim but not ERA5. Please check.

Thank you for this comment. The superseded version of CR2MET (version 1.0) was produced with data from ERA-Interim; however, the newest version (i.e., 2.0) incorporates data from ERA5. Unfortunately, there is no published work related to CR2MET version 2.0. To clarify this issue, we added the following information to the end of the paragraph describing the product (L127–129): *"These estimates are produced by combining rain gauge observations with reanalysis data from ERA5, while CR2MET version 1.0 of this product was produced using ERA-Interim data (Boisier et al., 2018)."*

R2C9: Lines 149-154: It will be interesting to report the control of elevation to the precipitation differences in the products. Representative West-East elevation and precipitation cross sections for three zones may help in this regard.

We agree with Reviewer 2 that it would be very interesting to report how elevation relates to P differences in the products. However, we believe that we would need many cross-sections to represent the gradient of P related to elevation across Chile due to the complex topography and climatology of the country. Instead, we complemented Figure 2 by adding Figure 3 (please see JPC1), which we believe manages to capture how P products vary over the different elevations.

R2C10: Lines 166-171: There are over 20 references in 6 lines. Reducing this number to 2-3 for each point the authors are targeting may help brevity. Same in line 225.

True! We have reduced the number of references (keeping what we considered to be the most relevant ones) in both sections.

R2C11: Figure 4: Caption: Replace "Leave-out-out" with "Leave-one-out."

Thank you for detecting this typo. It has been corrected in the new version of the article (please see Figure 6).

R2C12: Figure 5: The ECDF color-coding description should be corrected in the caption based on the legend (calibration-green, Ver-1 blue). I also suggest drawing the country outline thinner and with less intensity to improve visualization of the markers. If possible, a light gray shaded relief map will add elevation information.

Thanks for noticing this. We have corrected the colours in the caption and the outline of Chile has been coloured in light gray. Additionally, we added a very light hillshade; however, it is very faint because it must have high transparency to not interfere with the primary information of the plot. The updated figure (Figure 7 of the manuscript) is replicated below (Figure R19):

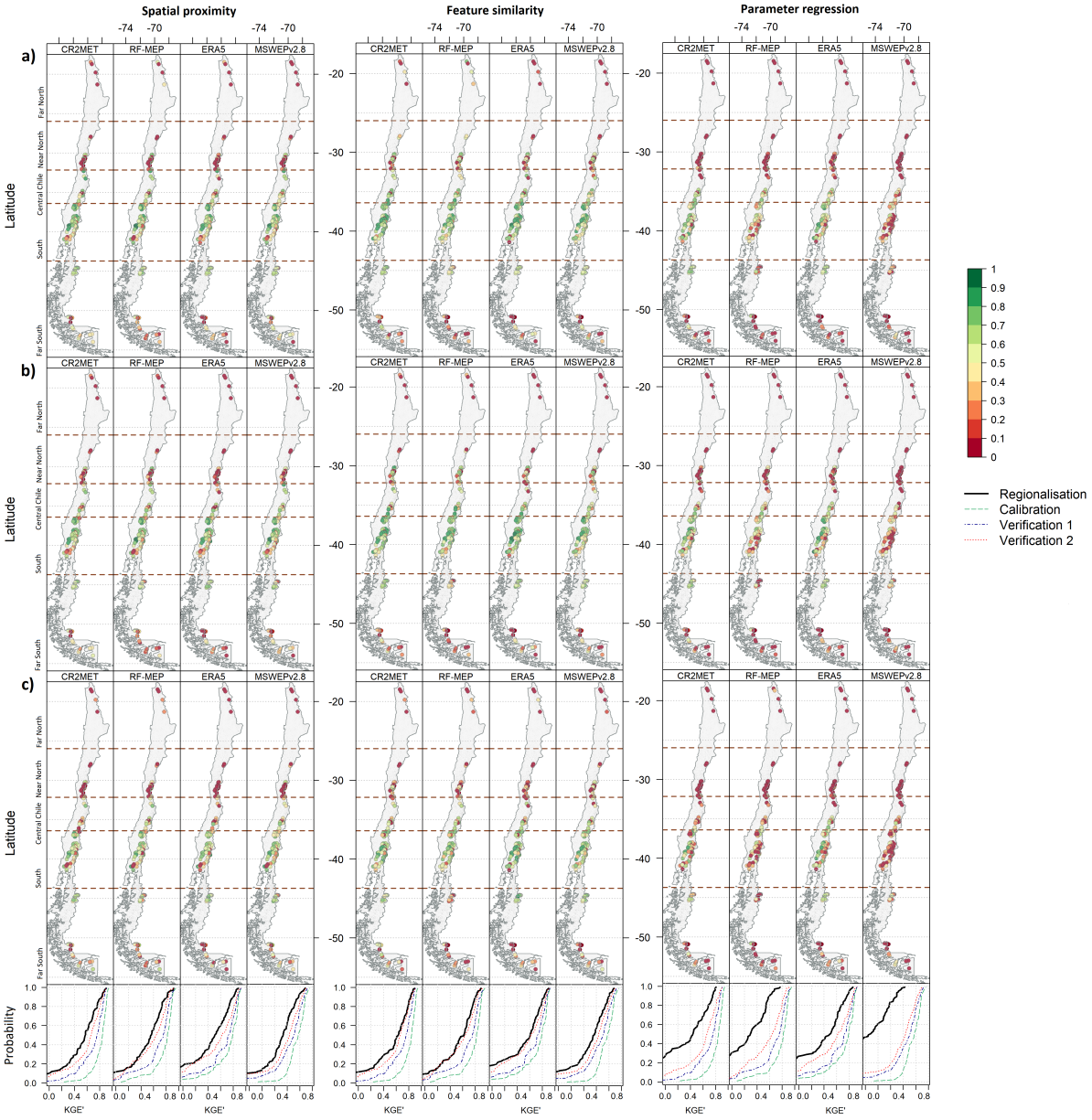

Figure R19. Spatial performance of the leave-one-out cross-validation results for the three regionalisation methods according to P product used to force TUWmodel. Results are presented for the: a) calibration (2000–2014); b)Verification 1 (1990–1999); and c) Verification 2 (2015–2018) periods. The panels beneath the map plots refer to the ECDFs of the corresponding regionalisation technique for the entire period of analysis (1990–2018) and P product (black) against the performances during the independent calibration (green), Verification 1 (blue), and Verification 2 (red) periods.

R2C13: Line 325: CR2METv2 – please use consistent acronym throughout the manuscript (earlier use was CR2MET).

Thanks! Now the product is referred to as "CR2MET" throughout the manuscript.

R2C14: Lines 335-337: "Choice of objective function does not impact the spatial performance of regionalization methods": This sentence seems to be an overstatement because different objective functions are sensitized to different components of the hydrograph (e.g log transformed flows will put more emphasis on low flows) and hence put more emphasis on fitting different model parameters. Instead of using a general statement, please use a statement valid for the present study. In Figure 5, comparison of ECDFs for P product columns between KGE' and AOF metrics indicates some differences. For example, compare feature similarity ECDF for ERA5 between KGE' and AOF metrics.

Thank you very much for these insightful comments! It would be very interesting to evaluate how the objective functions (in this case the KGE' and the AOF) fit the model parameters according to the emphasis they place on different components of the hydrograph. However, after evaluating our manuscript in response to reviewer comments (see our response to ET–General comments (Part 3) and ETC5), we have removed the AOF from the manuscript, with the intention of writing an additional paper introducing and analysing this new objective function in more detail. In this upcoming paper, we plan to include a detailed evaluation of how different objective functions affect the simulation of different components of the hydrograph.

R2C15: Lines 347-348: Please discuss which catchment characteristics are lacking that represent the hydrological behavior in this region.

Thank you for this interesting question. The following sentences have been added to the new version of the manuscript between L378–385: *The systematic lower performance of feature similarity compared to spatial proximity over the Far South (except for the case of ERA5) could be attributed to: i) the lack of catchment characteristics that represent the hydrological behaviour of this complex area dominated by polar and temperate climates; and ii) the low amount of potential donor catchments (eleven for latitudes $> 49°S$), combined with their varied hydrological regimes. For the most southern catchments, the highest P intensities occur during March–May, while the lowest P occurs between June–August, which differs to catchments throughout the rest of the country (Alvarez-Garreton et al., 2018, their Figure 9). This may affect the hydrological simulations when model parameters from catchments located $< 49°S$ are transferred to these far southern catchments."*

R2C16: Lines 354-355: "For feature similarity, all P products yielded a performance similar to that obtained for individual basin verification during the dry Verification 2 period". This statement somehow contradicts with the information provided in Figure 6. In Figure 5, ECDFs for regionalization and calibration/verification represent different time periods and hence their comparison may yield to misinterpretation. In Figure 6, however, regionalization and calibration/verification periods overlap and it can be seen that feature similarity performance is lower than Verification 2 performance.

Thank you for this important observation. In our response to JPC2, we acknowledge that the presentation of results over mixed time periods was confusing in our initial submission. Therefore, we have modified the manuscript to show the individual regionalisation results during calibration, Verification 1, and Verification 2. The figure showing the ECDFs (Figure 7 in the

updated manuscript) now shows the spatial distribution of regionalisation performance (i.e., the KGE' values) for each time period, as well as the ECDFs for these time periods, allowing it to be directly compared with Figures 4, 5, 6, 8, which all separate the results into these three time periods.

We thank Reviewer 2 for pointing out that the sentence between L354–355 was incorrect, and we have therefore deleted it from the revised manuscript. Additionally, we understand that the ECDFs for the regionalisation and calibration/verification periods overlap in Figure 7. However, we consider that by making the aforementioned changes to other figures in manuscript, we directly and comprehensively compare the performance between independent calibration and verification periods with the regionalisation performances over the corresponding periods. In this sense, rather than including another direct comparison according to period, we believe that we have added new information to the reader by comparing the regionalisation performance over the full period with the independent model performance over the near-normal and dry periods in the ECDFs. To ensure that the overlap in periods is clear to the reader, we added a clarification to the figure caption, which now reads: *"The panels beneath the map show the ECDFs of the corresponding regionalisation technique and P product (black) for the entire analysis period (1990–2018) compared with the performances during the individual catchment calibration (green), Verification 1 (blue), and Verification 2 (red) periods for the corresponding P product."*

R2C17: Line 376: Indicate that Figure 8 is for AOF.

Thank you for the comment. After removing the AOF from the revised manuscript, this sentence and the corresponding figure have been removed.

R2C18: Figure 9 and Lines 401-402: "The performance of parameter regression did not change substantially when evaluated with and without nested catchments." Actually, the performances with and without nested catchments look precisely the same (median values, distribution etc.). Is this related to using one RF model for each TUWmodel parameter? (see Lines 283-287). Please clarify here (Lines 400-401) and in Lines 283-287). I see some explanation in Lines 499-502, but more direct evidence should be provided.

Thank you for picking up on this point. We have adjusted the text between L431–433 and added a clarification about the performance of parameter regression. This sentence now reads: *"The change in performance of parameter regression was negligible after the exclusion of nested catchments because, in the particular case of Chile, excluding only a few catchments had a negligible effect on the non-linear relationships between model parameters and the selected climatic and physiographic characteristics (see Table 4)."*.

The explanation of this behaviour is not directly related to the use of RF in parameter regression, rather it can be attributed to the following reasons (included in the Discussion Section between L566–569: *"i) the degree of nestedness, as the unique geography of Chile limits, to some extent, the number of nested catchments within any larger catchment (only 10 of the 100 selected catchments contained more than three nested catchments); and ii) the percentage of catchments that are nested (42% in this study, compared to 65% in the Austrian case study)."*

R2C19: Lines 416-417: MSWEP is also a merged product as stated in Line 411? Please clarify.

Completely true! The comparison was intended to be with ERA5 instead of MSWEPv2.8. However, we have modified the respective paragraph as a result of the other comments and the modified sentence can be observed between (L464–468) and now reads: *"The decomposition of the KGE' also showed the ability of the TUWmodel to compensate for the total volume of P, as the β component was close to the optimum value, particularly for calibration and Verification 1 (see Appendix C), which can be attributed to the improved detection of P events of the merged products (regarding RF-MEP, see Baez-Villanueva et al., 2020). This can also be observed for MSWEPv2.8, as it produced the best performance over snow-dominated catchments under dry conditions (Verification 2)."*

R2C20: Lines 422-423: See earlier Comment on ERA5. Also, ERA5 reported 4 times more precipitation in the dry north (Figure 2) that may lead to low regionalization performance.

Thank you for this comment. Despite the model compensating to some degree for differences in P products (see our response to JPC3), it is true that the large disparity in total P compared to other products is a likely cause of the low performance in regionalisation. We have elaborated the text in the manuscript, which is now as follows (L471–478): *"The lower performance obtained in regionalisation with ERA5 in the Far North compared to the other P products (median values < 0.18 for feature similarity in all periods) can be attributed to its high P values, which are likely due to the lack of ground-based P stations over Chile in the development of the product. The incorporation of ground-based stations has the potential to: i) compensate for overestimations caused by the evaporation of hydrometeors before they reach the ground (Maggioni and Massari, 2018); and ii) improve event-based detection skills (Baez-Villanueva et al., 2020; Zhang et al., 2021). The latter is evident in CR2MET and MSWEPv2.8, which are both based in ERA5 but included several rain gauges in the Far North, and have a higher performance than ERA5 (see Figures 2, 3, and S1)."*

R2C21: Line 424: Please revise the language: "The use of ground observations as they:"

Thank you for noting the lack of clarity about how we linked this sentence to the previous one. We have revised the text (L471–474) to: *"The lower performance obtained in regionalisation with ERA5 in the Far North compared to the other P products (median values < 0.18 for feature similarity in all periods) can be attributed to its high P values, which are likely due to the lack of ground-based P stations over Chile in the development of the product. The incorporation of ground-based stations has the potential to: i) compensate for..."*

R2C22: Line 465: Please clarify "opposite" here. Does it refer to the regime?

Thank you for the comment. After removing the AOF from the manuscript, this sentence was also removed.

R2C23: Line 477: Please move the left parenthesis before the year of the reference.

Thank you for the comment. After removing the AOF from the manuscript, this sentence was also removed.

R2C24: Line 489: Please indicate "snowmelt" recession curves.

Thank you for the comment. After removing the AOF from the manuscript, this sentence was also removed.

R2C25: Line 494: Typo “Figure” 9.

Thank you for catching this error. We added the word *"Figure"* to the manuscript. Please see L548.

R2C26: Line 512: Remove “model” after TUWModel.

The word *"model"* has been removed from the revised manuscript. Please see L585.

R2C27: Boxplots: Please describe box limits and vertical line (median) in first boxplot figure caption.

Good point! Figure 4 now includes the following text: *"The solid line represents the median value, the edges of the boxes represent the first and third quartiles, and the whiskers extend to the most extreme data points within 1.5 times the interquartile range from the box. The blue line indicates the optimal value for the KGE'."*

R2C28: Lines 518-519: See earlier comment on ERA5 (merged product)

Thank you very much for this and the other comments. As explained in R2C4, we do not consider ERA5 to be a merged product.

Elena Toth (ET)

ET–General comments (Part 1): The manuscript presents the results of a certainly massive amount of work, aiming at analysing the performances of a set of gridded rainfall products when regionalising the parameters of a rainfall-runoff model, to be used at ungauged river sections in a data-scarce region. It addresses therefore a very interesting, and novel topic, with important practical impacts and potential for improving predictions in ungauged basins (PUB) in many regions of the world.

On the other hand, the topic is complex and the aim of the work is perhaps a bit too ambitious for a single paper.

Thank you for your insightful comments, which have helped us to substantially improve the quality of the manuscript. We appreciate that you find our work novel, interesting and ambitious. We acknowledge that the scope of the work is quite large and might be difficult to summarise into a single manuscript. That being said, our manuscript now has a more refined scope and we therefore believe that the readers would benefit from the inclusion of both independent catchment calibration/verification and regionalisation exercises within a single publication. To address the issue of complexity, we have removed some analyses from the manuscript and we believe that the article is now more concise and easier to follow. Please see below for the detailed responses to your comments.

ET–General comments (Part 2): Given the main final focus of the work, the paper requires first of all a more thorough analysis and comparison of the rainfall products, that should be carried out at catchment scale and over the different seasons (see, for example, Tarek et al, HESS 2020). The interpretation of such comparison should be based on the knowledge of the differences among the rainfall products, considering and explaining in particular the differences among their sources, first and foremost the availability and use of raingauges (that I personally, as a hydrologist, still find the most reliable information on actual rainfall). Such analysis is currently only partial in the paper and it is probably its weakest point.

Following such analysis, a more detailed presentation of the performances of the streamflow simulations when the model is "regularly" calibrated (not regionalised) may allow to understand if the products are indeed equally suitable to reproduce the rainfall fields over the catchment in a "reasonable" way, through the simulation of the rainfall-runoff transformation processes. I totally agree with the authors that the parameters may compensate for the differences in the rainfall fields, up to a given extent, and such analysis is therefore very interesting, especially for a data-scarce region. Some of the interpretations in the discussion and conclusions (see reference to Figure 5 in Sect 5.1) on the compensation are referred to the regionalisation results, while they should be demonstrated through the analysis of the results of the "calibrated" simulations.

Such first two steps would be enough for a paper of its own, without considering the regionalisation that is built on top of them. The importance of the rainfall estimates is certainly crucial also for model regionalisation purposes, but I think this is mainly due to an indirect impact of their reliability on the different climatic areas, extremely diversified in the Chilean region, and not directly to the regionalisation approaches: for this reason, a deeper understanding of such reliability is mandatory.

Thank you for raising these important points. Here, there are two aspects we strive to balance: to provide a clear and complete answer to our main objective within one single paper, while ensuring that the article remains focused. For this reason, we have decided to stick with one manuscript that addresses all points relevant to our central research questions, while removing interesting but not strictly necessary analyses that certainly can be part of a future publication.

We agree with your point regarding the benefit of including *"a more thorough analysis and comparison of the rainfall products"* prior to regionalisation. Please refer to our response to JPC1, where we describe in detail the additions made to the manuscript to address this concern. In addition to the changes described in our response to JPC1, we have also included the information regarding the number of rain gauges in Chile used by the P products to correct their estimates, which are 874 for CR2MET, 334 for RF-MEP, and 628 for MSWEPv2.8. This information has been added to L130, L133, and L152, respectively. Additionally, we now refer the reader to Figure S1 of the supplement (replicated below as Figure R20), which plots the locations of these rain gauges.

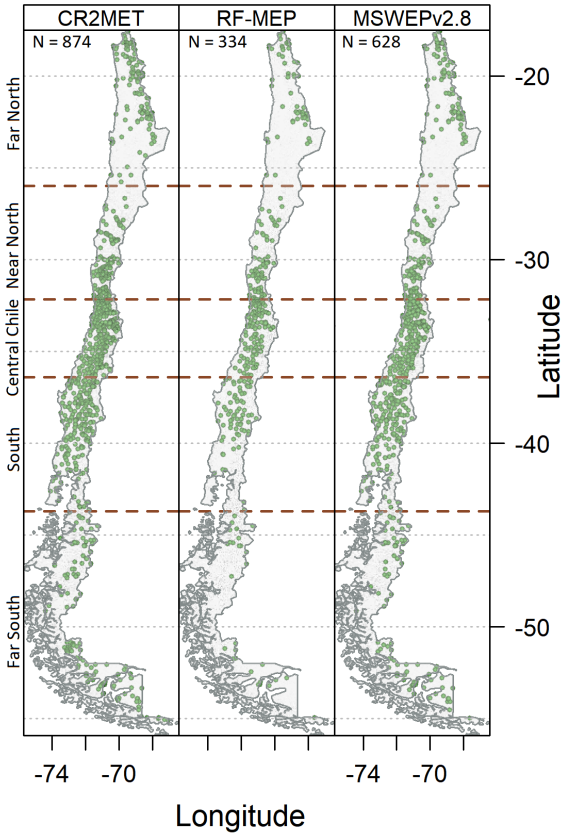

Figure R20. Rain gauges that each merged product used to construct their P estimates over Chile.

Regarding your point about the *"performance of the streamflow simulations when the model is regularly calibrated (not regionalised)"*, we also agree that a detailed discussion on this point adds value to the manuscript. Please refer to our response to JPC3, which goes into detail on how the TUWmodel parameters vary to compensate for differences between P products, as well as discussing monthly variations in the water balance components. Furthermore, to consider model performance in more detail prior to regionalisation, we have replaced Figure 3 in the original manuscript with Figure 5 in the new manuscript, which summarises model performance according to hydrological regimes (replicated as Figure 21 below). The results presented on this figure are discussed in detail in the manuscript in L325-345.

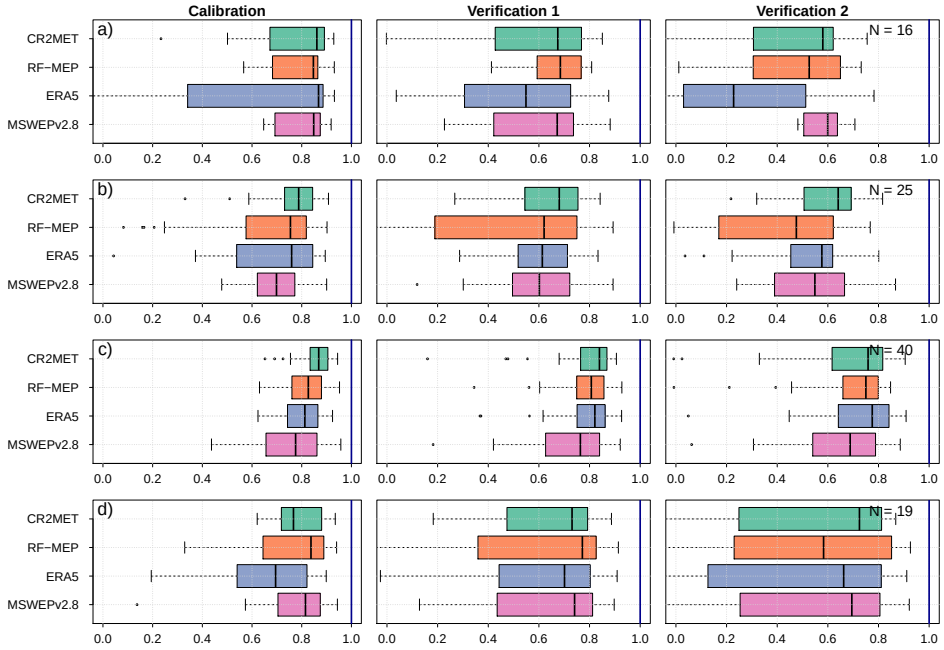

Figure 21. Performance of TUWmodel during calibration (2000–2014), Verification 1 (1990–1999) and Verification 2 (2015–2018), prior to any regionalisation over catchments with different hydrological regimes: a) snow-dominated; b) Nivo-pluvial; c) pluvio-nival; and d) rain-dominated.

Regarding your statement about the role of the reliability of P estimates, we agree that the selection of a reliable P product is crucial for hydrological applications, which motivated us to select only the best P products for our specific case study of Chile. Numerous studies have evaluated the performance of P products using rain gauges as ground truth. However, in data-scarce settings, the low density of rain gauge networks prevents us from having a sound benchmark that can be used to assess the reliability of the individual P products to force hydrological simulations, especially over high-elevations and complex topography. In this manuscript we demonstrated that the calibration of the hydrological model is able to compensate, to some extent, for differences in the volume, spatial distribution, and occurrence of P. Therefore, for the purposes of answering our

central research objectives, we consider an assessment of the reliability of the selected P products to be outside of the scope of the manuscript.

We strongly believe that the updated version of the manuscript is more comprehensive in its analyses and discussion, and thus conveys a stronger and more complete message to the readers. Therefore, we believe that it is worth making the effort to keep both the hydrological modelling for individual calibration and verification as well as for regionalisation in one article rather than separating it into two or more papers. This is primarily because we need to include both aspects in order to comprehensively answer our research objective of analysing how the choice of P products affects the results for different regionalisation techniques.

ET–General comments (Part 3): In an already very complex framework, the authors have added a second, complicated objective function (AOF), which would be justifiable only if the study focussed much more on the analysis of the simulation of the different parts of the hydrograph (low flows, flow duration curves, baseflow...), which is not instead the focus of the present work, and the issue is only partially addressed, and only late, in the discussion section on the hydrological signatures. As it is, the second OF lengthens and complicates the paper, and I would suggest to remove it.

This is a very valid comment. Although we were pleased with the results we have shown by using the AOF, we understand that it adds an extra level of complexity to the manuscript that is not needed to answer the specific research question of the paper (i.e., how the selection of P products affects the regionalisation performance). Therefore, we have removed the AOF from the updated manuscript with the intention to write a second paper focusing on introducing and analysing this objective function in detail.

ET–General comments (Part 4): Another analysis that I think may be removed or restructured is the comparison of the regionalisation performances over the different hydrological regimes. This is very interesting issue to study per se, but i) the rationale for the separation of the basins into such groups ("visual screening"?) has not been explained (and the hydrological regimes seem to overlap only partially with the identified climatic regions, which I find a bit puzzling) and ii) the results of section 4.2 are not particularly significant: the rainfall products do not show any clear pattern and the differences are probably more due to the more pronounced rainfall errors or difficulties of the model in reproducing the streamflow over the different regimes, independently of the regionalisation procedure. In order for such section to be useful, it should be preceded by an analysis - over the catchments belonging to such regimes - of the reliability of the rainfall maps and of the model performances when the model is not regionalised but calibrated (other analyses to be added in the first two steps. . .)

Firstly, thank you for highlighting that a detailed description related to catchment classification according to hydrological regimes was missing. The following explanation was added to L112–118: *"We adjusted the classification of these catchments according to hydrological regime, building on the classifications presented in several national and regional technical reports (e.g., DGA, 1998, 1999, 2004a, b, c, 2006, 2016a, b, 2018) by visually analysing the contribution of solid and liquid P to the mean monthly Q. These regimes were classified as: i) snow-dominated, ii) nivo-pluvial, i.e., snow-dominated with a rain*

component, iii) pluvio-nival, i.e, rain-dominated with a snow component, and iv) rain-dominated, as shown in Figure 1d. Figure A1 shows conceptual hydrographs for each of these regimes and is presented in Appendix A." This conceptual figure is replicated below (Figure 22).

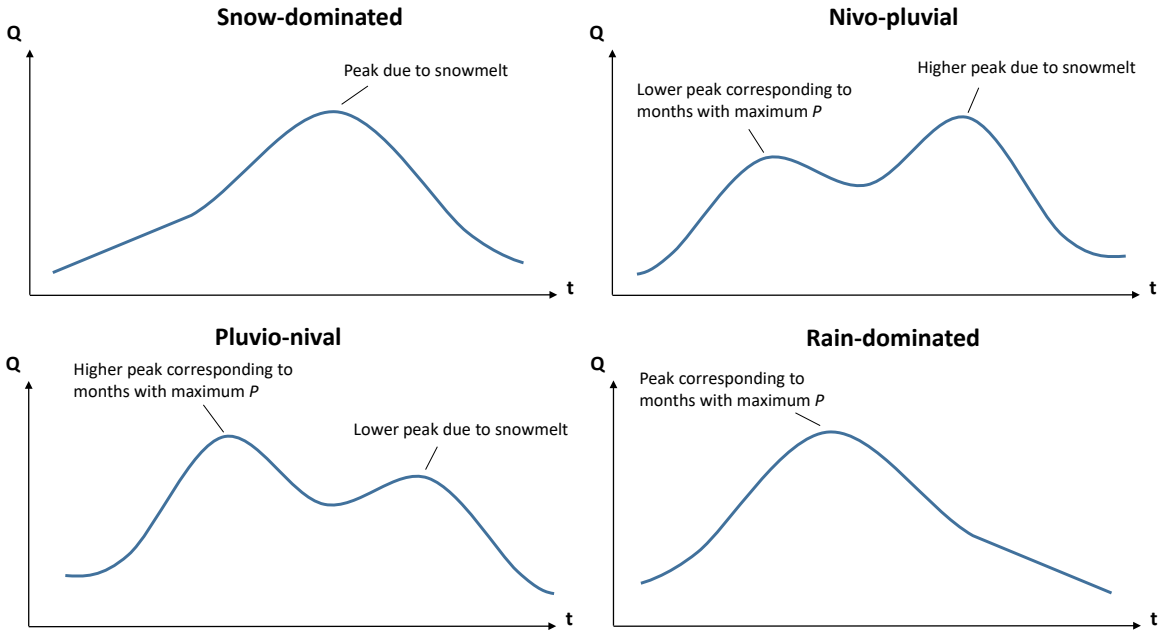

Figure 22. Conceptual illustration of the hydrological regimes used to classify the 100 near-natural catchments used in this study.

Next, you mention that the identified hydrological regimes overlap only partially with the climatic regions. Although the hydrological regimes implicitly consider the climatic conditions of the respective catchments, they do not always coincide with a particular Köppen-Geiger climate classification (see Figure 1c and 1d of the manuscript), because the hydrological response of most catchments is strongly shaped by multiple climates, which in turn are linked to the complex topography of Chile. For this reason, the classification of catchments by hydrological regimes makes sense in the context of Chile, despite not matching entirely the climatic zones. In this sense, the validity of this classification is demonstrated in the marked differences in the distribution of model parameters for different regimes, which is evident in Figure 12 of the manuscript and Figures S12–S14 of the supplement.

We have identified and discussed how the model parameters values can compensate for different P inputs to match observed runoff under different hydrological regimes (please see our response to JPC3). Building on this information, we strongly believe that the results for regionalisation performance are now presented in a more valuable way to the reader, due to the extra

information provided about individual model response across different hydrological regimes. Furthermore, having updated Figure 5 (performance during the independent calibration and verification periods according to hydrological regime) of the manuscript (here, Figure R21), the results shown in this Figure are now directly comparable to Figure 9, which allows us to disentangle the reasons for model performance during individual calibration and regionalisation.

ETC1: ll 68-74 and Section 3.1.1: As I wrote above, the main limitation of the study is the lack of a detailed description of the rainfall products: the main difference is in their sources, whereas it is not a matter of being gridded and probably also the spatial scale is less relevant than how much they rely on ground measurements. The ERA5 is probably the product with less dependence on raingages, since reanalyses assimilate a number of both measured and remotely sensed information within the numerical models, but mainly atmospheric and ocean measures and not ground data, I think. But all the other products you use are, on the basis of your description, based on a merging of reanalysis (ERA5) and ground-based data (CR2MET and RF-MEP), and for MSWEP including also use of satellite data. For all such products some information on the location and temporal coverage of the raingages data is needed, in order to understand their differences. It is written that RF-MEP uses 331 gauges, but not if CR2MET uses the same data. And MSWEP is based on 77000 gauges globally, but how many are in Chile? (and are they the same used in the other products?). Adding a map with the locations of such gauges would also help to interpret the reliability of the products over the different parts of the country.

Thank you for this comment. We agree that the comparison between the selected P products needed to be improved. We now show the location of the ground-based measurements that were used in the development of the three merged products (i.e., CR2MET, RF-MEP, and MSWEPv2.8) in Figure S1 of the supplement, and we added information to L130, 133, and 152 about the number of rain gauges used to correct estimates, as described in the response to ET-General comments (Part 2).

ETC2: In addition, it would be useful to know if one of these products (CR2MET?) is considered, or it has been demonstrated in previous studies, to be more reliable or if it is the 'reference' product for the Chilean meteorological or hydrological offices (which one is included as meteorological driver into CAMELS-CL?)

CR2MET was developed specifically for continental Chile and uses all the available rain gauges across the country, hence it is considered as the 'reference' P product by the Chilean meteorological and hydrological agencies. This has been added to the manuscript between L129–130: *"As CR2MET was developed specifically for continental Chile and uses all the available rain gauges (874 across Chile; see Figure S1 in the supplement), it is considered as the 'reference' P product of Chile"*.

Furthermore, CR2MET is used as a meteorological driver in CAMELS-CL together with CHIRPSv2, MSWEPv1.1 and TRMM 3B42v7. However, as we did not use P-related data from CAMELS-CL, this information was not added to the manuscript as we believe that the addition of such information could lead to confusion.

ETC3: As written above, I would suggest adding more information on the comparison of the rainfall fields, that should be carried not only over the mean annual values (Figure 2), but over the different years (so to differentiate also the "near normal", and "dry" periods cited when identifying the verification periods in Sect 3.3), over the different seasons, and, especially

important, at catchment scale, in order to preserve the consistency among the rainfall forcing and the corresponding streamflow (see, for example, Tarek et al, HESS 2020). For each product, it may be shown, for example the yearly time-series of the box-plots representing the Mean Areal Precipitation for the catchments)

Thank you for this comment. We agree that a thorough comparison between the selected P products was missing in the initial version of our manuscript. Please see our response to JPC1, which describes our efforts to address all your comments on this topic by providing a detailed comparison of: i) the differences between the P products for the entire period as well as for the near-normal and dry conditions; ii) seasonality of the products at the catchment scale by macroclimatic zone for the entire period as well as for the near-normal and dry conditions; and iii) spatial distribution of the length of dry and wet spells, as well as extreme P values. Following your idea, we included boxplots representing mean monthly catchment-averaged P amounts (see Figure 2d).

ETC4: Section 3.3: the fact that the first 10 years (1990-1999) are used both as warm-up period and as verification period is a bit confusing: ten years of warm-up period is not necessary, and in this way there seems to be no warm-up for Verification 1 period? At least one year of warm-up would be needed... This point may be clarified.

Thank you for bringing up this important point. We agree that 10 years of warm up might be excessive for most catchments, however, we selected 10 years to be conservative in the initialisation of the model stores over catchments with varied climates and geomorphological characteristics. For the case of Verification 1 (1990–1999), we used the calibrated model parameters to ran TUWmodel from 1983 (where all P products were available), to include a warm-up period that was as long as possible. We elaborated the sentence describing the warm-up in the manuscript (L218–219), which now reads as follows: *"For calibration purposes, we used the first ten years as a conservative warm-up period to initialise the model stores, as in Beck et al. (2020)"*. In addition, we clarified the warm-up period used for Verification 1 (L223–224): *"To initialise model stores for the Verification 1 period, we used an 8-year warm up period due to P product availability."*

ETC5: pag. 11: as written above, I would remove the AOF objective function.

Thank you for this suggestion. Although the inclusion of the AOF is very interesting (and we were able to demonstrate an improved performance over the arid catchments), following your comments we decided to remove it from the manuscript. We plan to work on an additional manuscript to present the results obtained with the AOF, which we believe are relevant and very interesting.

ETC6: Section 4.1.1: more information should be given on the performances of the models when not regionalised: first of all distinguishing the performances over the different climatic regions and interpreting the efficiencies in relation to the possible lack of reliability of the rainfall forcing in specific catchments/regions. In addition, looking at the entire boxes and whiskers (and not at the median values only) in Figure 3, I do not agree that ERA5 is equally (or better) performing than the other products: perhaps more details on local performances may help to better understand.

Thank you for this comment. We have made an effort throughout the manuscript to address these points as follows:

- We have made an effort throughout the manuscript to describe the results considering the dispersion of the products and not only the median values;
- We have separated the entire period of analysis into calibration, Verification 1 and Verification 2 periods to analyse the results under near-normal and dry conditions (please see JPC2) and we now include the analysis of the hydrological regimes for the independent calibration/verification (i.e., not regionalised); and
- We included an evaluation of how the calibration of TUWmodel compensated for the differences between P products according to hydrological regime (please see JPC3). With this evaluation we demonstrated how the model compensated, to some extent, for the increased P amounts shown by ERA5, producing overall comparable results to the merged products.

We acknowledge that it would be very interesting to analyse the performance of the catchments according to climate; however, due to the climatic and physiographic complexity of Chile it is challenging to attribute a single climate to a specific catchment. Instead, we introduced the five major macroclimatic zones of Chile (described in Zambrano-Bigiarini et al., 2017) to assess the differences in P products according to localised regions (please see JPC1). As mentioned in our response to ET-General comments (Part 4), we consider that the classification into hydrological regimes makes sense in the context of Chile, and more importantly, to answer our specific research questions.

As the idea of separating the behaviour of the catchments per climate is very interesting, we computed the distribution of the calibrated model parameters per macroclimatic region of Chile (Figure 1). These results are shown below in Figures R23–R27, which show that the distribution of model parameters according to hydrological regime is similar to those of the macroclimatic zones in regions where a single hydrological regime is dominant. For example, the parameters of catchments located over the South region are in agreement with those of the pluvio-nival catchments. However, in the Far South there are catchments with varied regimes, and as a consequence, the model parameters differ from those classified by hydrological regime. Note that for the Far North there are only four catchments for each boxplot.

Even though the previous figures are interesting, we decided not to include them in the final version of the manuscript, as we believe that: i) their inclusion will add another degree of complexity to the manuscript; ii) these results are partially in agreement with those obtained in the evaluation according to hydrological regime; and iii) it would be more difficult to attribute the compensation mechanisms of TUWmodel if the catchments included in the analysis do not exhibit similar hydrological behaviour.

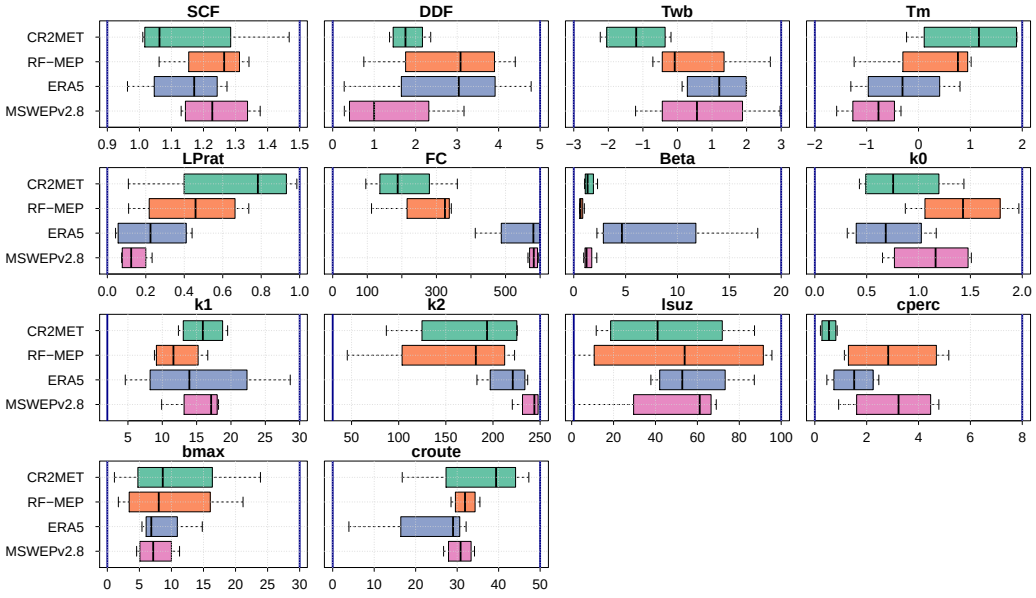

Figure R23. Model parameters obtained through calibration of catchments located in the Far North region. The vertical blue lines indicate the upper and lower limits of the parameter ranges.

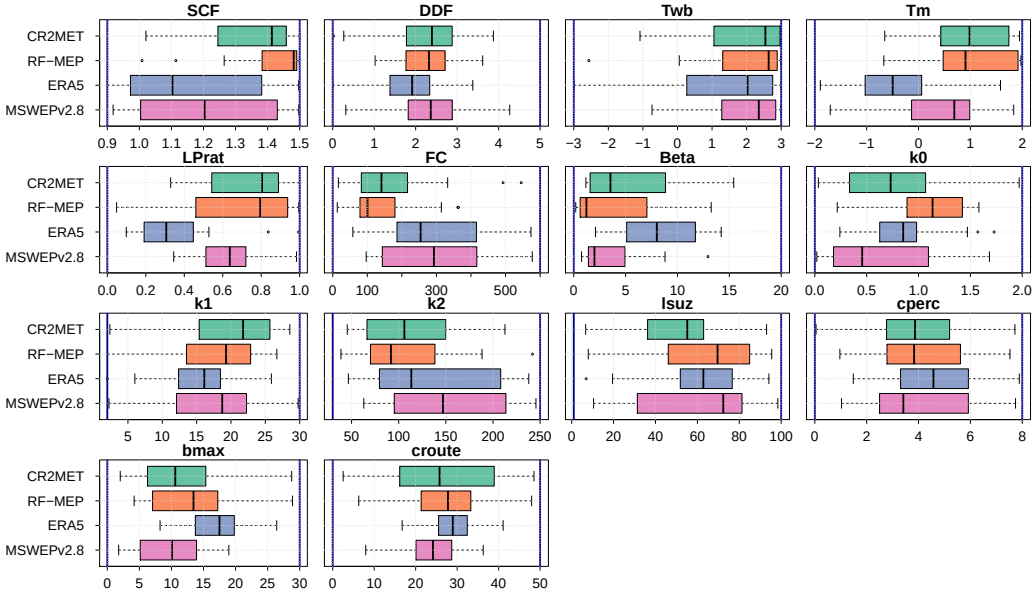

Figure R24. Model parameters obtained through calibration of catchments located in the Near North region. The vertical blue lines indicate the upper and lower limits of the parameter ranges.

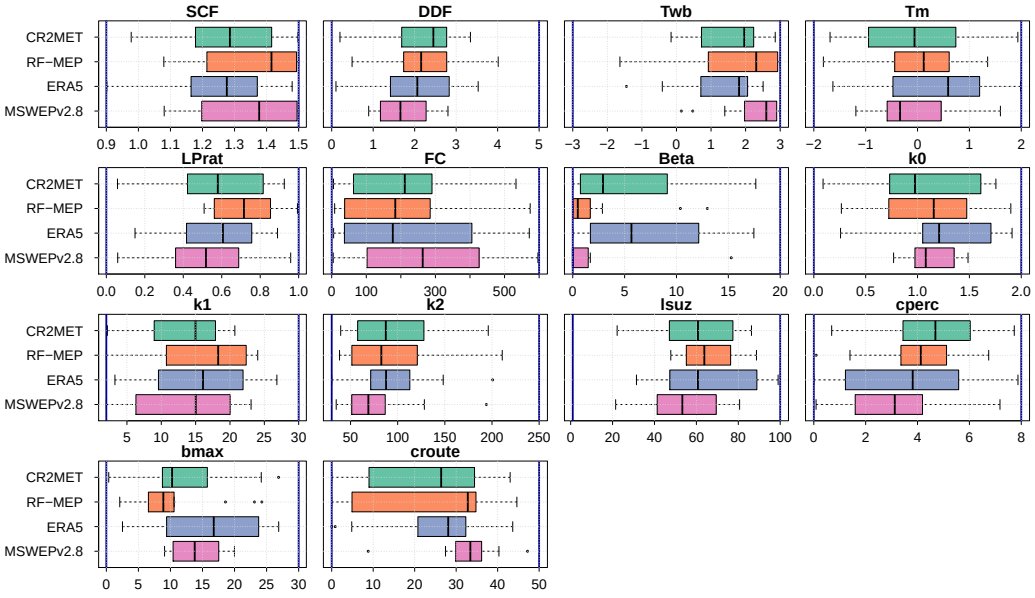

Figure R25. Model parameters obtained through calibration of catchments located in Central Chile. The vertical blue lines indicate the upper and lower limits of the parameter ranges.

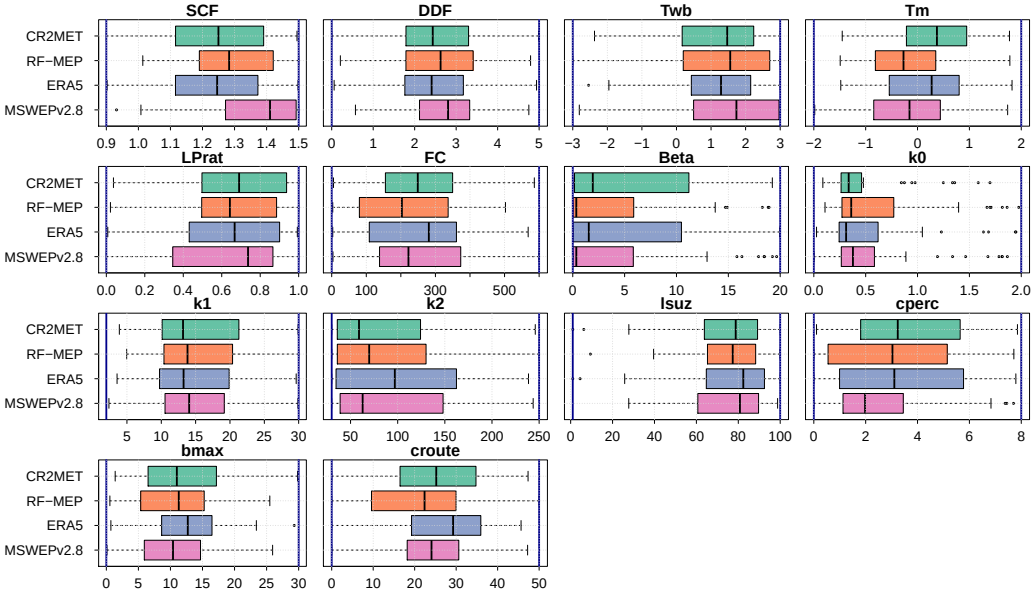

Figure R26. Model parameters obtained through calibration of catchments located in the South region. The vertical blue lines indicate the upper and lower limits of the parameter ranges.

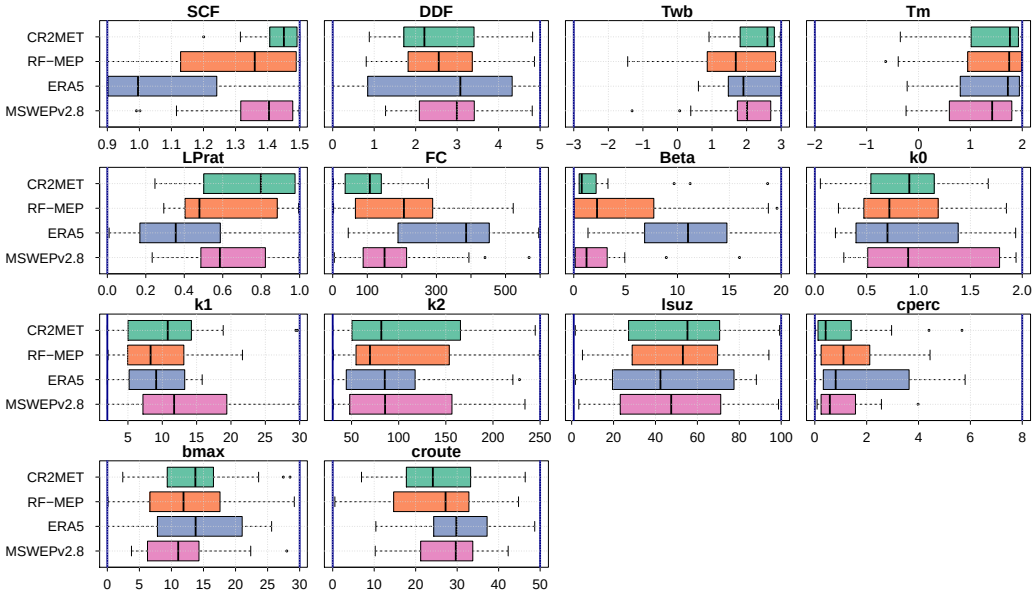

Figure R27. Model parameters obtained through calibration of catchments located in the Far South region. The vertical blue lines indicate the upper and lower limits of the parameter ranges.

ETC7: Section 4.2.2: as above written, if this section is preserved, there should be a similar analysis for the 'at-site parameterised', not regionalised simulations (and more information on how the regimes have been identified).

This is completely true, such extra information is needed for the section on hydrologic regimes to be comprehensive and valuable to the reader. As mentioned in our response to ET–General comments (Part 4), we decided to keep the hydrological regimes in the article because their inclusion brings forward interesting aspects related to the performance of regionalisation techniques. Please refer to this aforementioned response where we cover this comment in detail, as well as the response to JPC3.

ETC8: l.1: I suggest to clarify in the abstract that what is regionalised are the parameters of a rainfall-runoff model.

Thank you for this suggestion. We emphasised in the abstract that we regionalised the parameters of a rainfall-runoff model. The updated sentence is as follows (L7–9): *"We assessed the ability of these regionalisation techniques to transfer the parameters of a rainfall-runoff model using different P products, implementing a leave-one-out cross-validation procedure."*

ETC9: Figure 2: I would use a linear scale for the legenda, since the current colours do not clearly distinguish the regional differences in the rainfall values among the products.

We appreciate this suggestion. However, after trying a linear scale (and many other scales), the current colour palette was selected in order to emphasise the spatial differences between P products at both low and high ranges of P. We believe that the use of a linear colour scale would cause a loss of comparability between P products for low and medium P values.

ETC10: l.323: I do not agree that the spatial proximity and feature similarity results are so close: adding an x-axis grid in Figure 4 would highlight that Feature similarity performs better, looking at the entire boxes.

Thank you for highlighting how we can better improve the comparability of our results in our figures. An x-axis grid has been added and we changed the wording in (L353) from "quite close" to "relatively close".

ETC11: ll 340-344: actually, not only RF-MEP but all products (a part from ERA5) are a merging of reanalysis and ground-based observations, from what I understood.

Thank you for this observation. Due to the reorganisation of the manuscript, this sentence not longer appears in the revised version of the manuscript. In addition, see our response to comments R2C4 and ET–General comments (Part 2).

ETC12: lines 349-354 (and bottom panel of Figure 5) are more related to the results presented in Section 4.2.1 (and in Figure 6) and may be moved there?

Thank you for this observation. The results between L349–354 of the previous manuscript related to the ECDFs in Figure 7 (Figure 5 of the previous version of the manuscript) have been moved to Section 4.2.1 (L401–406).

ETC13: Section 4.2.1: maybe at least a first comment on the differences among the rainfall products shown by Fig, 6 should be added/moved here.

Thank you for the suggestion! We have commented on the differences in how the model performed using different P products between L392–395: *"There are marked differences in performance according to the P product used to force the TUWmodel, regardless of the regionalisation method and the evaluated period. For example, ERA5 has more dispersion in the KGE' values compared to other products for the cases of feature similarity and spatial proximity; while for parameter regression, it tends to perform the best."*

References

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., and Ayala, A.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset, Hydrology and Earth System Sciences, 22, 5817–5846, https://doi.org/10.5194/hess-22-5817-2018, 2018.

Baez-Villanueva, O. M., Zambrano-Bigiarini, M., Beck, H. E., McNamara, I., Ribbe, L., Nauditt, A., Birkel, C., Verbist, K., Giraldo-Osorio, J. D., and Thinh, N. X.: RF-MEP: A novel Random Forest method for merging gridded precipitation products and ground-based measurements, Remote Sensing of Environment, 239, 111 606, https://doi.org/10.1016/j.rse.2019.111606, 2020.

Bao, Z., Zhang, J., Liu, J., Fu, G., Wang, G., He, R., Yan, X., Jin, J., and Liu, H.: Comparison of regionalization approaches based on regression and similarity for predictions in ungauged catchments under multiple hydro-climatic conditions, Journal of Hydrology, 466, 37–46, 2012.

Beck, H. E., van Dijk, A. I., De Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, Water Resources Research, 52, 3599–3622, 2016.

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., Van Dijk, A. I., Huffman, G. J., Adler, R. F., and Wood, E. F.: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS, Hydrology and Earth System Sciences, 23, 207–224, https://doi.org/10.5194/hess-23-207-2019, 2019.

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I., and Wood, E. F.: Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments, Journal of Geophysical Research: Atmospheres, 125, e2019JD031 485, 2020.

Boisier, J. P., Alvarez-Garretón, C., Cepeda, J., Osses, A., Vásquez, N., and Rondanelli, R.: CR2MET: A high-resolution precipitation and temperature dataset for hydroclimatic research in Chile, EGUGA, p. 19739, 2018.

Clerc, M.: From theory to practice in particle swarm optimization, in: Handbook of Swarm Intelligence, pp. 3–36, Springer, 2011a.

Clerc, M.: Standard particle swarm optimisation from 2006 to 2011, Particle Swarm Central, 253, 2011b.

DGA: Plan director para la gestión de los recursos hídricos en la cuenca del río San José, Tech. rep., Dirección General de Aguas, Santiago, https://snia.mop.gob.cl/sad/ADM600v1.pdf, 1998.

DGA: Recursos hídricos compartidos con la República Argentina : ficha temática de la cuenca del río Grande de Tierra del Fuego, Tech. rep., Dirección General de Aguas, Santiago, https://snia.mop.gob.cl/sad/CUH2087.pdf, 1999.

DGA: Cuenca Quebrada de Tarapacá, Tech. rep., Dirección General de Aguas, Santiago, https://mma.gob.cl/wp-content/uploads/2017/12/Tarapaca.pdf, 2004a.

DGA: Cuenca Río Loa, Tech. rep., Dirección General de Aguas, Santiago, https://mma.gob.cl/wp-content/uploads/2017/12/Loa.pdf, 2004b.

DGA: Cuenca del Río Elqui, Tech. rep., Dirección General de Aguas, Santiago, https://mma.gob.cl/wp-content/uploads/2017/12/Elqui.pdf%0A, 2004c.

DGA: Evaluación de los recursos hídricos superficales de las cuencas de los ríos Petorca y La Ligua V Región, Tech. rep., Dirección General de Aguas, Santiago, https://snia.mop.gob.cl/sad/SUP4496.pdf, 2006.

DGA: Análisis integral de soluciones a la escasez hídrica, región de Arica y Parinacota : informe final, Tech. rep., Dirección General de Aguas, Santiago, https://snia.mop.gob.cl/sad/REH5720.pdf, 2016a.

DGA: Actualización de Información y Modelación Hidrológica Acuíferos de la XII Región, de Magallanes y la Antártica Chilena : Informe definitivo etapa II, Tech. rep., Dirección General de Aguas, Santiago, https://snia.mop.gob.cl/sad/SUB5698.pdf, 2016b.

DGA: Herramientas de gestión y actualización de los modelos numéricos del acuífero de Copiapó : informe final, Tech. rep., Dirección General de Aguas, Santiago, https://snia.mop.gob.cl/sad/SUB5851v1.pdf, 2018.

Garambois, P.-A., Roux, H., Larnier, K., Labat, D., and Dartus, D.: Parameter regionalization for a process-oriented distributed model dedicated to flash floods, Journal of Hydrology, 525, 383–399, 2015.

Karl, T. R., Nicholls, N., and Ghazi, A.: Clivar/GCOS/WMO workshop on indices and indicators for climate extremes workshop summary, in: Weather and climate extremes, pp. 3–7, Springer, 1999.

Maggioni, V. and Massari, C.: On the performance of satellite precipitation products in riverine flood modeling: A review, Journal of Hydrology, 558, 214–224, 2018.

McIntyre, N., Lee, H., Wheater, H., Young, A., and Wagener, T.: Ensemble predictions of runoff in ungauged catchments, Water Resources Research, 41, 2005.

Neri, M., Parajka, J., and Toth, E.: Importance of the informative content in the study area when regionalising rainfall-runoff model parameters: the role of nested catchments and gauging station density, Hydrology and Earth System Sciences, 24, 5149–5171, https://doi.org/10.5194/hess-24-5149-2020, 2020.

Oudin, L., Andréassian, V., Perrin, C., Michel, C., and Le Moine, N.: Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, Water Resources Research, 44, 2008.

Parajka, J., Merz, R., and Blöschl, G.: A comparison of regionalisation methods for catchment model parameters, 2005.

Széles, B., Parajka, J., Hogan, P., Silasari, R., Pavlin, L., Strauss, P., and Blöschl, G.: The added value of different data types for calibrating and testing a hydrologic model in a small catchment, Water Resources Research, p. e2019WR026153, 2020.

Zambrano-Bigiarini, M. and Baez-Villanueva, O.: Tutorial for using hydroPSO to calibrate TUWmodel, https://doi.org/10.5281/zenodo.3772176, https://doi.org/10.5281/zenodo.3772176, 2020.

Zambrano-Bigiarini, M. and Rojas, R.: A model-independent Particle Swarm Optimisation software for model calibration, Environmental Modelling & Software, 43, 5–25, https://doi.org/10.1016/j.envsoft.2013.01.004, 2013.

Zambrano-Bigiarini, M., Nauditt, A., Birkel, C., Verbist, K., and Ribbe, L.: Temporal and spatial evaluation of satellite-based rainfall estimates across the complex topographical and climatic gradients of Chile, Hydrology and Earth System Sciences, 21, 1295, 2017.

Zelelew, M. B. and Alfredsen, K.: Transferability of hydrological model parameter spaces in the estimation of runoff in ungauged catchments, Hydrological Sciences Journal, 59, 1470–1490, 2014.

Zhang, L., Li, X., Zheng, D., Zhang, K., Ma, Q., Zhao, Y., and Ge, Y.: Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach, Journal of Hydrology, 594, 125 969, https://doi.org/10.1016/j.jhydrol.2021.125969, 2021.