Replies to review

"Towards robust seasonal streamflow forecasts in mountainous catchments: impact of calibration metric selection in hydrological modeling"

Diego Araya, Pablo A. Mendoza, Eduardo Muñoz-Castro and James McPhee

We thank the reviewer for their time in commenting on our paper. We provide responses to each individual point below. For clarity, comments are given in italics, and our responses are given in plain blue text.

Reviewer #2

The authors have addressed most of my comments. The manuscript is easier to read and the main conclusions are better highlighted by the new figures. I have one last major comment that I would like the authors to respond to, and a couple of minor comments.

Major comment:

I would like to thank the authors for responding to my comments regarding the distinction between calibration and evaluation periods for the analysis of hindcast performance. I realise that the main idea of my comments may not have been entirely clear, and I apologise for that. Although I agree that using the longest available period is important to obtain reliable hindcasts, I still think that there needs to be a distinction between evaluation and calibration periods for the hindcast performance analysis, as is done for the hydrological consistency analysis. As I understand it, one of the aims of this study is to evaluate the differences in hindcast performance and hydrological consistency for different objective functions and initialisation times. So, the main point is not just to get good hindcast performance - which in fact would require the use of some sort of threshold to determine what is acceptable performance - but to compare hindcast performance between different modelling setups. So I think the question is whether you would still get the same rankings in terms of the best objective function for hindcast performance if you had distinguished the evaluation from the calibration periods. For example, I would like to know if the VE-sep objective function is still the best objective function in terms of CRPSS when looking at the evaluation period. One solution to determine whether this changes the results in terms of objective function ranking could be to produce a very simple figure for the TUW model, 1 September CRPSS and the four objective functions of Figure 9, but with the distinction between evaluation and calibration periods. This could be added to Section 5.5 and would either replace or complement Figure 11.

We appreciate the reviewer's comment, since it enriches our discussion on sample size issues in hindcast verification, especially when developing and testing seasonal forecasting methods. To address this observation, we examined the sensitivity of the CRPSS for September 1 hindcasts, to the stratification of the full verification sample (i.e., 33 WYs) between hydrologic model calibration (April/1994 – March/2013; i.e., 19 WYs) and evaluation (April/1987 – March/1994 and April/2013 – March/2020; i.e., 14 WYs) datasets (Figure 12). Here, we used parameters calibrated with the five representative OFs and the TUW model, using data from the period April/1994 – March/2013. The results show that the VE-Sep remains the top-performing objective function in terms of CRPSS, while Split KGE yields the worst results. Further, the rankings of the other objective functions (NSE, ModKGE, and KGE(Q)+NSE(log(Q))) vary depending on the verification period, and CRPSS values are higher during the calibration period compared to the evaluation datasets. We have added this new figure (to complement Figure 11) and text (L445-451) in Section 5.5, following the reviewer's recommendation.

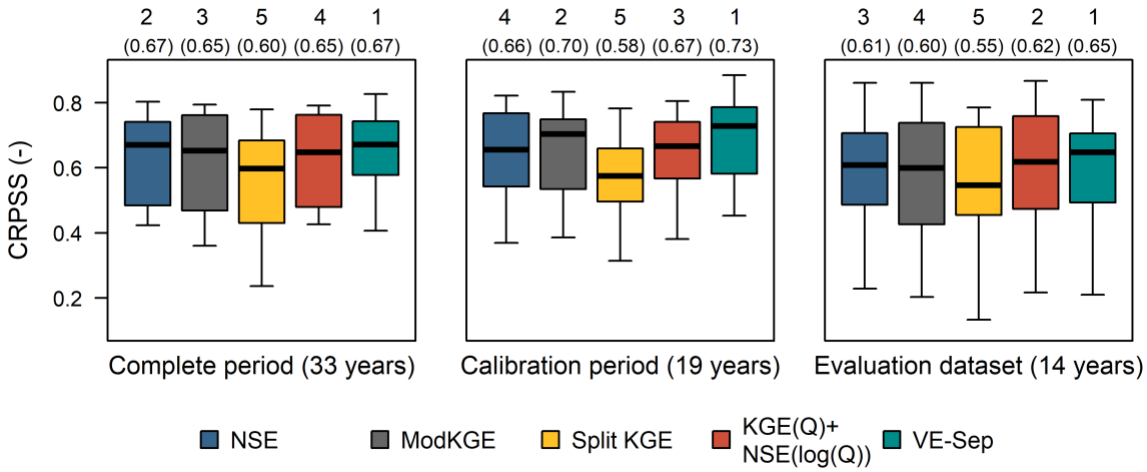

Figure 12. Comparison of CRPSS for September 1 hindcasts obtained with the five representative objective functions and the TUW model. Each panel contains results for a different hindcast verification period: (left) 33 WYs (full period); (middle) 19 WYs (calibration period); and (right) 14 WYs (i.e., evaluation dataset April/1987 – March/1994 and April/2013 – March/2020). Each boxplot comprises results from the 22 case study basins and one objective function. The boxes correspond to the interquartile range (IQR, i.e., 25th and 75th percentiles), the horizontal line in each box is the median, and whiskers extend to the ±1.5·IQR of the ensemble. The numbers in parentheses denote the median CRPSS among all basins, and the numbers above the OF ranking based on that median, being 1 the best.

Other comments:
Why were elevation bands defined only for CemaNeige?

We did not use snow elevation bands in SAC/SNOW 17 and TUW models because preliminary experiments showed that the benefits of adding these on the KGE of daily flows were marginal. We stress that the use of three models does not seek to provide comparisons among different model structures; instead, we aim to examine to what degree our results and conclusions can be model-dependent. We have made these points in L182-186:

"While the CemaNeige is configured with 10 elevation bands, the snow routines of TUW and SAC-SMA (i.e., SNOW-17) are implemented in a lumped fashion because preliminary experiments with these models showed that the benefits of adding snow bands on the KGE of daily flows were marginal. We stress that the use of three models does not seek to provide comparisons among different model structures; instead, we aim to examine to what degree our results and conclusions can be model-dependent."

Although I did not mention it in the first report, I find Figure 10 difficult to understand. In order to better support the main conclusions related to Section 4.4, I suggest that the authors reduce the content of this figure. For example, only the results for two initialisation times and TUWmodel could be shown in this figure.

We have modified this Figure to show the results for the TUW model, following the reviewer's recommendation. We have decided to keep the five initialization times to illustrate the progression of the relationship between hindcast performance and catchment attributes, and the results for the SAC and GR4J models are included in the supplements (Figure S12). Accordingly, we have made some re-organization and modifications on the text (L342-355):

"We now explore the factors that control seasonal hindcast quality, and the extent to which the choice of calibration metric impacts the connections inferred from our sample of catchments. Figure displays results for the TUW model only, and the full results (including GR4J and SAC) are available in the Supplement. In general, the choice of calibration metric affects more the strength, rather than the sign,

of the relationships between hindcast quality and catchment attributes. In particular, we find that the correlations between CRPSS and catchment descriptors obtained with Split KGE (which maximizes hydrologic consistency), are weaker than those obtained with other calibration metrics (e.g., see results for baseflow index with TUW, interannual runoff variability with all models, and fraction of precipitation falling as snow with all models).

We find statistically significant correlations between CRPSS and the baseflow index ($\rho \sim 0.2 - 0.8$) with the three models, being ModKGE ($\rho = 0.49$), VE-Sep ($\rho = 0.70$), and VE-Sep ($\rho = 0.41$) the objective functions that maximize such relationship for September 1 when using TUW (Figure), GR4J and SAC (Figure S12), respectively. Figure shows significant correlations between CRPSS and the interannual variability of runoff ($\rho \sim 0.0 - 0.6$) – especially for September 1 hindcasts ($\rho = 0.53$ for VE-Sep/TUW, $\rho = 0.64$ for ModKGE/GR4J and $\rho = 0.62$ for VE-Sep/SAC). Also positive, but generally weaker correlations are obtained between hindcast skill and p-seasonality ($\rho \sim -0.6 - 0.0$), as well as the fraction of precipitation falling as snow ($\rho \sim 0.0 - 0.4$)."

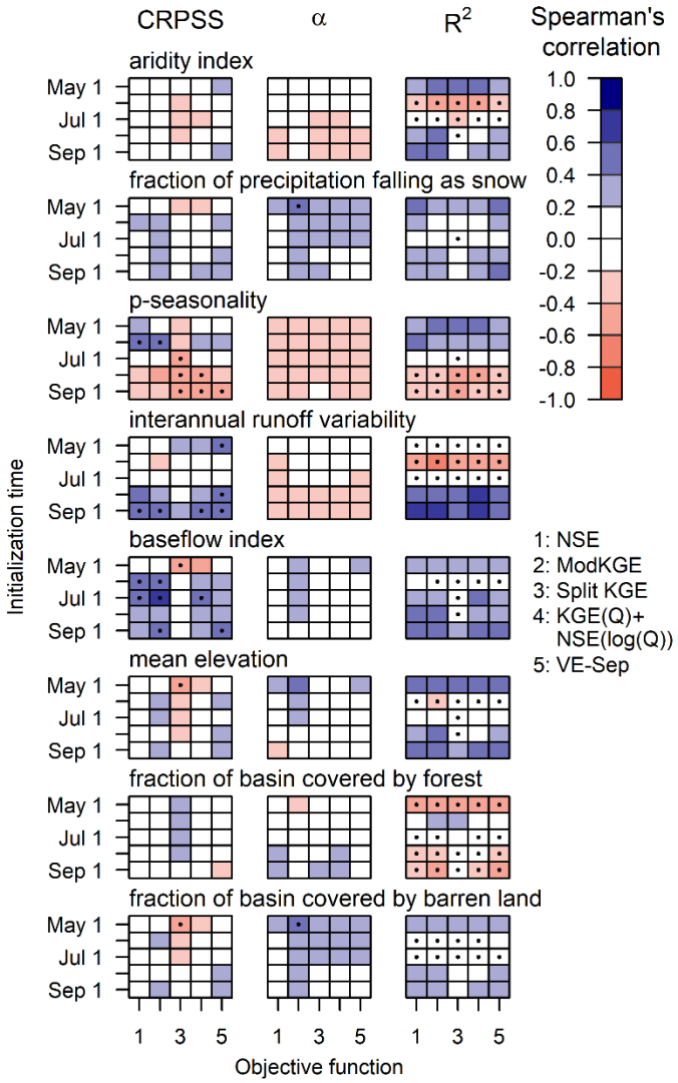

Figure 10. Spearman's rank correlation coefficients between catchment characteristics (shown in different rows) and the CRPSS (left), α reliability index (center), and the coefficient of determination R2 (right) obtained for seasonal streamflow hindcasts (period April/1987 – March/2020), produced with the five representative objective functions (x-axis in each color matrix), different initialization times (y-axis in each color matrix), and the TUW model. Black dots indicate statistically significant ($p < 0.05$) correlations.

Out of curiosity, I looked at your scripts and saw that you tested GR5J and GR6J. Was there a reason for keeping only GR4J? Does using GR6J solve the problem with the X4 parameter (unit hydrograph

time constant which is often close to the upper limit for almost half of the catchments after calibration, probably when baseflow is high)?

Preliminary experiments with the GR5J and GR6J models (Figure R1) – reported by Araya (2022) – provided worse results in some basins compared to GR4J using the KGE as objective function, especially during the period 2013-2020 (Figure R1c). Even though GR6J addresses the issue with the X4 parameter (whose upper limit is 20) in some basins, it still provides X4 values near the upper limit in some catchments. In fact, X4 > 19 with the GR6J (GR4J) model in 9 (7) out of 22 basins.

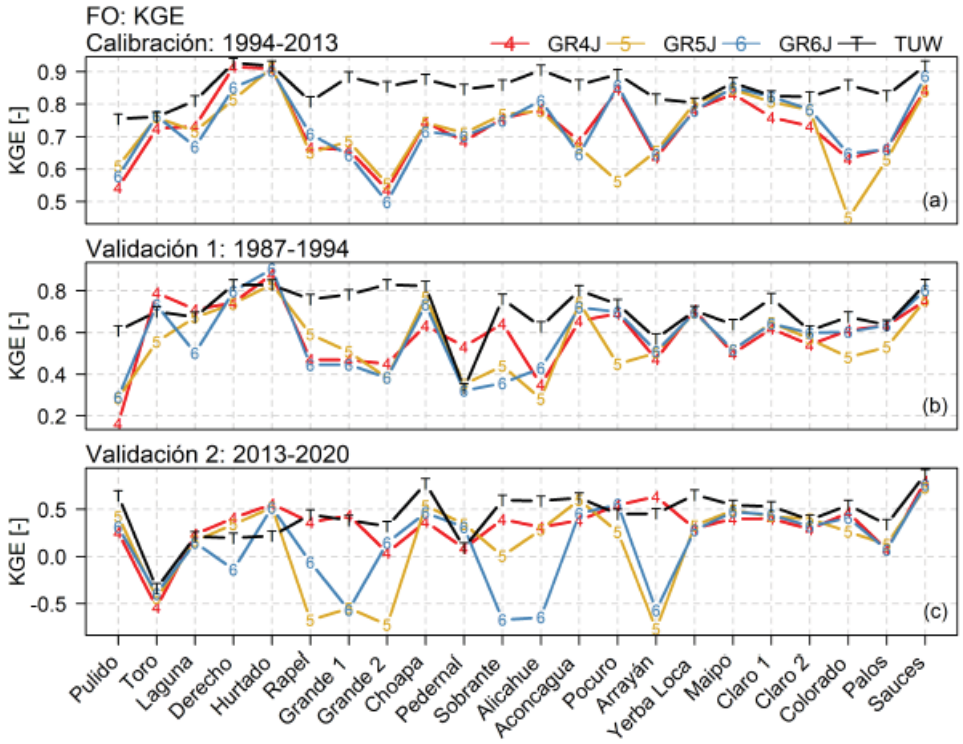

Figure R1. KGE computed with daily flows for all basins, obtained from calibrations with GR4J, GR5J, GR6J (all including CemaNeige) and TUW models using KGE as the objective function. The results are displayed for the periods (a) 1994-2013, (b) 1987-1994, and (c) 2013-2020.

Table R1. Parameter values obtained from calibrating GR4J and GR6J (both including CemaNeige), using KGE as the objective function.

BNA Basin	X_4 GR4J	X_4 GR6J	BNA Basin	X_4 GR4J	X_4 GR6J
3414001	14.45	19.49	5200001	19.91	12.74
4302001	15.24	5.29	5410002	20.00	20.00
4301002	15.58	19.97	5411001	1.03	1.03
4311001	20.00	16.26	5722001	20.00	15.34
4501001	8.30	3.32	5721001	20.00	19.99
4522002	1.62	1.61	5710001	17.72	19.89
4511002	20.00	20.00	6027001	1.14	1.51
4513001	1.60	19.53	7103001	1.02	1.36
4703002	20.00	20.00	7112001	1.04	1.31
5101001	1.00	1.13	7115001	1.40	1.44
5100001	1.63	19.47	8104001	1.00	1.30

L197: why does the model evaluation also cover part of the spin-up period between 1987 and 1994? Perhaps I have missed an explanation somewhere.

We have modified the text to clarify this point. Regarding the spin-up period use in model calibration (L189-191):
"To compute the calibration objective function, we use modeled and observed streamflow data from the period April/1994 – March/2013 because it spans a diverse range of hydroclimatic conditions, considering the period April/1986 – March/1994 for model spin-up."

Regarding the spin-up period used to produce runoff simulations for the evaluation dataset (L197-201):
"Model evaluation is conducted by computing performance metrics with data from two periods: (i) April/1987 - March/1994, which is hydroclimatically diverse, and (ii) April/2013 – March/2020, which is characterized by unprecedented and temporally persistent dry conditions (Garreaud et al., 2017, 2019). To produce runoff simulations for each period, the preceding eight years (i.e., April/1979 - March/1987 and April/2005 – March/2013) were used for model spin-up."

Figure 5: increase size by rearranging panels.

We have increased the panel size by reducing the horizontal scale of the figure and, and we have adjusted the y-axes scale to enlarge the boxes. Additionally, we have compressed the panels to use the space more efficiently.

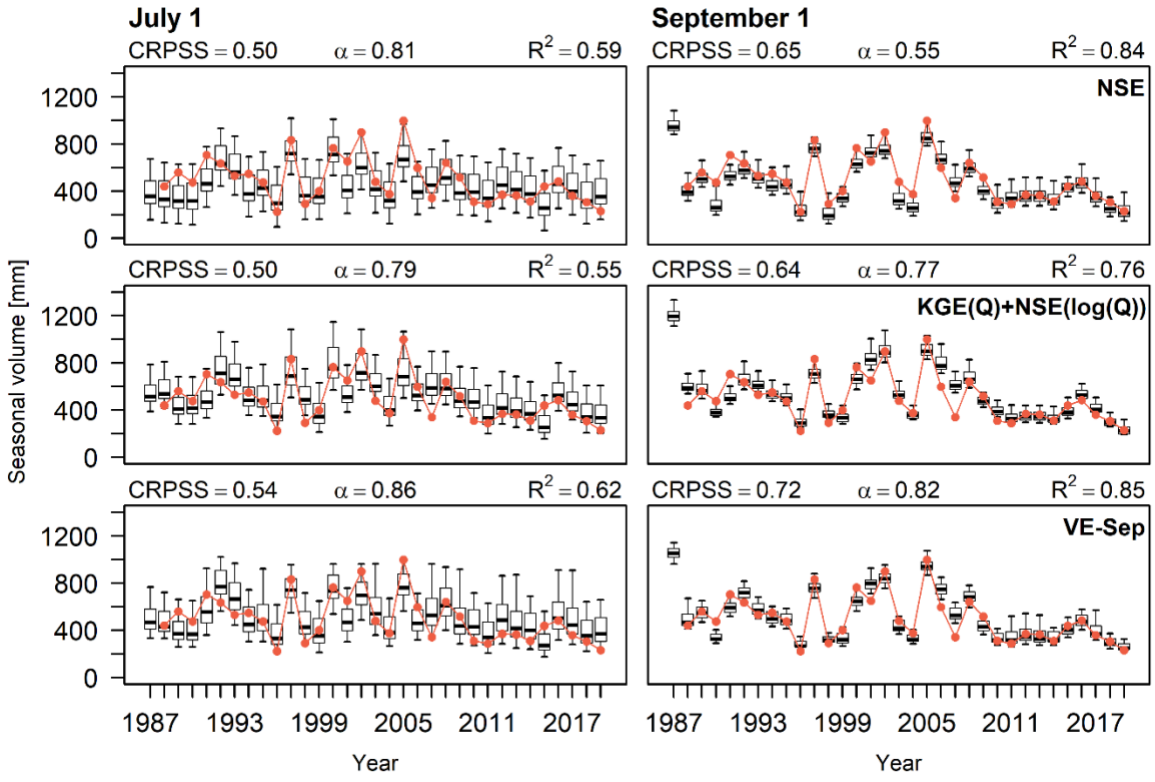

Figure 5. Time series with ESP seasonal hindcasts (i.e., September-March runoff) initialized on July 1 (left panels), and September 1 (right panels) for the Maipo at El Manzano basin. The boxes correspond to the interquartile range (IQR, i.e., 25th and 75th percentiles); the horizontal line in each box is the median, whiskers extend to the ±1.5·IQR of the ensemble, and the red dots represent the observations. The results were produced with the TUW model, using parameters obtained from calibrations conducted with NSE, KGE(Q)+NSE(log(Q)) and VE-Sep (see details in Section 3.1). Each panel displays the CRPSS, the reliability index α, and the coefficient of determination R2 (computed using the ensemble hindcast median).

Figure 9: use the same scale for the y-axis. Why is "Split KGE" in second place for TUW? The median of the average bias in hydrological signatures seems to be lower than the median of KGE+NSE(log). Have you ranked the objective functions after rounding up the variable?

We now use the same scale for the y-axis of the panels that belong to the same row.

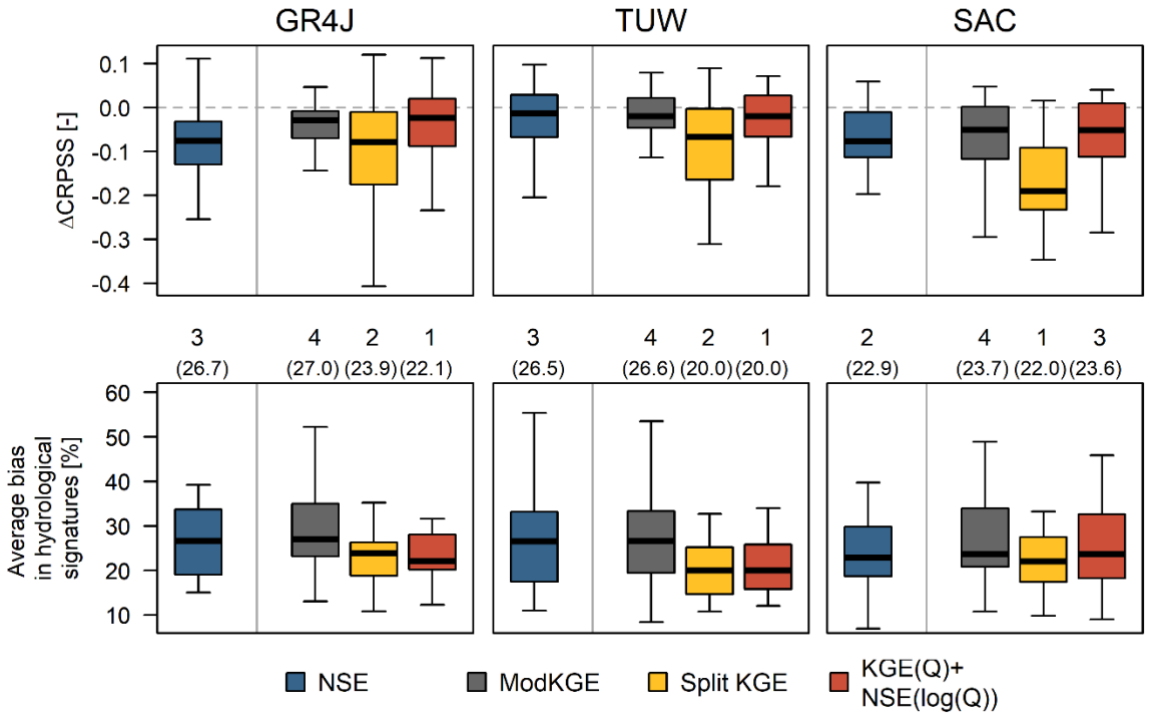

Figure 9. Variations in September 1 CRPSS due to the choice of popular and alternative objective functions (shown in different boxplots), relative to the best performing OF in terms of forecast quality (VE-Sep). The dashed line indicates no difference (i.e., no loss) in forecast performance. The bottom panel display the average bias in hydrological signatures (computed over the calibration and evaluation periods) with the associated ranking (being 1 the best in terms of hydrological consistency), and median average bias obtained from the sample of basins (in parentheses).

Split KGE is in second place for TUW because its median is 20.004%, while the median for KGE+NSE(log) is 19.988%. As expected, such small difference is very hard to see, even after zooming in (see attached figure). Note that the metrics were not rounded up to rank the objective functions.

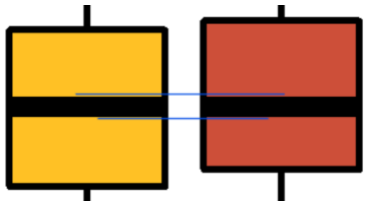

Figure S1: why is the period covered here only between 2014 and 2016?

Figure S1 was created to complement the results of Figure 4, which shows simulated and observed daily hydrographs for the period April/2014 - March/2016, and runoff seasonality for the evaluation dataset (April/1987 – March/1994 and April/2013 – March/2020) at the Maipo at El Manzano River basin, along with performance metrics. In the manuscript, it is stated, 'Similar results are obtained [...] for the remaining basins…' which is supported by Figure S1, which shows the daily KGE and the coefficient of determination between simulated and observed mean monthly runoff, for all 22 basins and the same periods examined in Figure 4. Hence, Figure S1 displays daily KGE for the period April/2014 - March/2016 to maintain consistency with Figure 4.

We have modified the caption of Figure S1 to clarify this:
“Kling-Gupta Efficiency (KGE) between simulated and observed daily streamflow for the period

April/2014-March/2016, using parameter values obtained with different calibration objective functions and hydrological models (upper panel); and coefficient of determination (R^2) between mean monthly simulated and observed runoff averages for the evaluation dataset (April/1987 – March/1994 and April/2013 – March/2020, from evaluation period (bottom panel). Each boxplot comprises results from the 22 case study basins. The boxes correspond to the interquartile range (IQR, i.e., 25th and 75th percentiles), the horizontal line in each box is the median, and the whiskers extend to the ±1.5·IQR of the ensemble. The points correspond to outliers beyond the whiskers' range."

References

Araya, D. (2022). Evaluación de la metodología ESP para la generación de pronósticos de caudales de deshielo en cuencas de Chile Central (in Spanish). Available at: https://repositorio.uchile.cl/handle/2250/185501