Author response

We thank the editor and referees for their careful reading and helpful comments. Our reply is given below. The page and line numbers correspond to the modifications done on the revised manuscript.

Reviewer 1

Line 23, “actual skill” here sounds ambiguous since there is no other information explaining this term, which might lead to misunderstanding. → We have replaced the term ‘actual skill’ with ‘overall skill’ with additional information explaining this term. (Page 1. Line 25)

Line 25, please add brief information on the methods that you use to get the conclusion that precipitation is the most important variable. → We have added explanations of the methods. (Page 1. Line 23-24)

Line 57, this is the first time that ESP is mentioned (excluding abstract), therefore full explanation is needed here. → We have added the full explanation of ESP. (Page 2. Line 58-59)

Line 77, to my knowledge, the reference Pechlivanidis et al., 2020 is not using ESP in the analysis, therefore cannot support the argument here. → We removed the reference from the paragraph. (Page 2. Line 77)

Line 93, I’m a little bit suspicious on this sentence here that “only a few studies” have used SEAS5 for seasonal hydrological skill assessment. For example, the reference you mentioned before Pechlivanidis 2020 is actually using SEAS5 at higher spatial resolution. → We wanted to emphasize that there are not many previous studies using ECMWF SEA5 and analysing the performance of SFFs compared to ESP. To clarify our intention, we modified the sentence. (Page 2. Line 93-95)

Line 139, what is the criteria of dividing the four seasons, are they based on precipitation or flow? → The criteria of dividing the four seasons is based on monthly precipitation. This is intended to maintain continuity with our previous research (Lee et al., 2023) and is consistent with the general seasonal classification in South Korea. (Page 4. Line 141-143)

Line 142, the information of annual variability is not shown in the figure but only in the text, right? → Yes. We included additional explanation and a reference (Lee et al., 2023) to support this. (Page 4. Line 145-146)

Line 143, typhoon and monsoon might not need to start with an uppercase character here. → It has been replaced to lowercase. (Page 4. Line 146)

Line 149, the abbreviation of KMA should be noted in the previous sentence when it is firstly mentioned. → We have added the abbreviation of KMA. (Page 4. Line 152)

Line 169, regarding SEAS5 data, here the period 1993-2020 is mentioned, but in the method part and in Figure 2, based on my understanding, the forecast period is 2011 to 2020. Please clearly specify this. → Our analysis focuses on the period from 2011 to 2020. However, we also analysed SEAS5 data from 1993 to 2010 to compute the bias correction factors. A detailed explanation of this process is provided as a Figure in the supplementary material (Figure S1). (Page 4. Line 173-174)

Line 181, SFFs has been mentioned many times already. → We have removed the full form of SFFs. (Page 5. Line 181)

Line 183, here CRPS is referred to as skill but later it is referred to as score (Line 258). → The issue of terminology for score and skill has been modified across revised manuscripts.

Line 188, the plot needs to be improved. To calculate CRPS needs the forecast (either ESP or SEAS5) and the reference (either real or pseudo-observation), therefore the arrows should lead from corresponding systems to the box of CRPS. However, this is not systematically shown in the plot. → We have improved Figure 2 to clarify our methodology and the term. (Page 5. Line 189)

Line 190, to my knowledge there is SEAS5 forecasts with higher spatial resolution that is available. → To clarify this issue, we have conducted a test to compare the forecasts with higher and lower resolution in three catchments. We have included additional explanation in Section 4.2 (Page 15. Line 544-551) and added the comparison result in the supplementary material (Figure S10).

Line 205, a potential problem for linear scaling on precipitation is, it might generate very large values. Have you had any solutions to avoid this? → In our study, we could not find any problem generating very large values. We also added references supporting our choice. (Page 6. Line 215-216)

Line 247, as defined in Eq.4? → We have corrected this typo. (Page 7. Line 262)

Line 265, what does SPFs stand for? Or maybe you mean SFFs? Otherwise please add the full name for the abbreviation. → We have corrected this typo. (Page 7. Line 280)

Line 270 and Line 258, redundant information. → We have removed the redundant sentence. (Page 8. Line 286)

Line 275, Major does not need an uppercase here. → We have removed the wording.

Line 275, here the CRPS of ESP is calculated using real observation as reference, it is correct? → Yes, it is. This is now clearly shown in Figure 2. (Page 5. Line 189)

Line 285, here comes the explanation of SPFs, but it is already mentioned many times before this. → SPF is a typo, we have amended this across the manuscript.

Line 310, here I would strongly recommend to distinguish skill from score, since you have CRPSS later which are actually skills, but here these are scores. → We have modified the term 'skill' to 'score' in those sentences where the CRPS is used.

Line 327, this part should be described in method session, and more details are needed for fully understanding. → We have moved this part to method section and added descriptions. (Page 5. Line 194-201)

Line 498, are these conclusions from Figure 8? Considering there are only two dry years and two wet years, the conclusion needs to be drawn carefully, otherwise it's not very scientifically valid. → We agree that having only two dry and two wet years means that we cannot draw definitive conclusions. We have revised our discussion to recognise this in our conclusions. Additionally, we have produced an additional Figure with the same analysis as Figure 8 but including data from the calibration period used to calculate bias correction factors. This extended the analysis to 5 dry years (1994, 2001, 2008, 2015, 2017) and 5 wet years (1998, 2001, 2002, 2011, 2020), respectively. (Please note that, due to the lack of observed data, here we can only use 7 catchments: Soyanaggang, Chungju, Andong, Imha, Hapcheon, Namgang, Sumjingang).

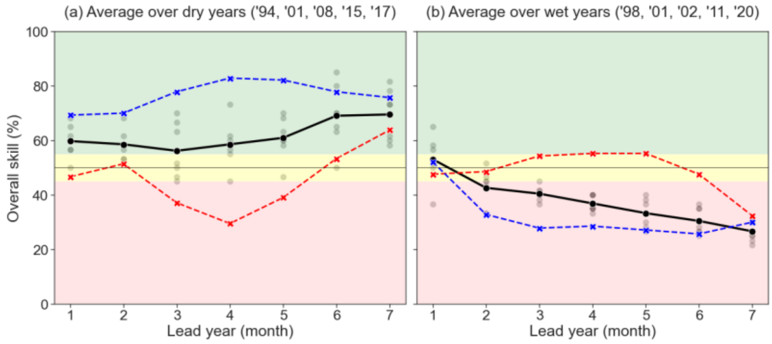

Figure R1. Overall skill of bias corrected SFFs over 7 catchments averaged over (a) dry years (mean annual P < 900mm) and (b) wet years (mean annual P > 1500mm) during all seasons (black lines), dry seasons (red dashed lines) and wet seasons (blue dashed lines). Here, mean annual precipitation is averaged across the catchments and years.

As shown in Figure R1, the results are generally consistent with Figure 8 (b, c), which is encouraging and, considering the available seasonal forecasts dataset (1993-), likely the broadest analysis that we can conduct. We also included additional information on this issue in the Discussion section. (Page 15. Line 529-531)

Figure S2, please explain which benchmark is used here to calculate from CRPS to skill (skill-P, skill-T).→ No benchmark was used here. To clarify this, we have modified Figure S2 (supplementary material).

Reviewer 2

L14: Replace “to link” with “to generate” → We have replaced it. (Page 1. Line 14)

L15-16 “at finer scales such as catchment”: A word seems to be missing → We have corrected it. (Page 1. Line 15-16)

L16: “generating SFFs (…) remains challenging” → We have corrected it. (Page 1. Line 15)

L19: “at catchment scale” → We have corrected it. (Page 1. Line 19)

L20: “over the last decade” → We have corrected it. (Page 1. Line 20)

L23 “actual skill”: this term is not clear at this stage. Please explicit.

→ We agree with you. To clarify our methodology and goal, we have replaced the term ‘actual skill’ with ‘overall skill’ in the abstract and add explanations on the ‘overall skill’. (Page 1. Line 25)

L24-25: this sentence states that you compare the skill of forecasts with the ESP. It seems odd. It is rather the ESP which is used as benchmark in the skill computation, and the comparison is carried via the calculation of the skill. Please clarify. → We have clarified this sentence. (Page 1. Line 25)

L30-31: are these “openly available”? → Yes, it is. We used ‘open-source Python package’. (Page 1. Line 32)

L57: this is the first occurrence of ESP, please explicit the term. → We have added the full explanation of ESP. (Page 2. Line 58-59)

L58: “by forcing a hydrological model with historical meteorological observations” → We have corrected this. (Page 2. Line 60)

L67-68: “Some of these studies focused” → We have corrected this typo. (Page 2. Line 68)

L95: “did not analyse” → We have corrected it. (Page 2. Line 95)

L108: “may be considered” → We have corrected it. (Page 2. Line 108)

L110 “on assessing the actual skill and comparing it with ESP”: I assume it is the actual skill of SFFs. This sentence may not work with the definition of the skill: either you compute the skill of SFFs with respect to a benchmark, and compute the skill of ESP with respect to that same benchmark, and then compare both skills, or you directly use the skill for the comparison (its intended use) and choose the CRPS of SFFs and ESP as numerator and denominator of the skill respectively. → We have corrected the sentence. (Page 2. Line 110)

Table 1: It would be informative to add Tmin and Tmax to this table, especially given that you have

catchments with snow which you later discuss. → We have added Tmin and Tmax in Table 1. (Page 4. Line 137)

(d,e,f): Here, instead of showing the average over all catchments, it would be interesting to represent the variability between catchments as it will later inform the variability in forecast skill.
→ We have added box plots in Figure 1 (d, e, f). (Page 3. Line 128)

In addition, in the caption: L133 "Mean monthly": isn't it rather the sum in the case of precipitation, PET and flow? → We have revised this error. (Page 3. Line 129-133)

L135 "variability of each weather variable": "of each weather and hydrological variable". Is it the inter-catchment variability or the inter-annual variability? → To clarify this issue, we have modified the caption. (Page 3. Line 129-133)

L149: Please introduce the abbreviation KMA here. → We have added full explanation of KMA. (Page 4. Line 152)

L156-160: Was the streamflow data generation done as a first step of this work? Do you make a distinction between "streamflow" and "flow" (L156)? After reading this paragraph, I was unsure whether you derived flow values (assuming flow and streamflow refer to the same variable) from measurements of river levels and a rating curve, or from a reservoir water balance, knowing measurements of reservoir levels, inputs other than inflows, and outflows, and then a rating curve. Is it because it is the second option being carried out that reservoir evaporation is mentioned? If so, are you improving on K-water's method? → We have changed the term 'streamflow' to 'flow' and clarify this point. (Page 4. Line 160)

L166: Please refer to:Johnson, S. J., and Coauthors, 2019: SEAS5: the new ECMWF seasonal forecast system. Geoscientific Model Development, 12, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019. → We have added the reference in that sentence. (Page 4. Line 169)

L174: Did you compute PET based on the Penman-Monteith method as mentioned L151, or did you retrieve PET forecasts directly from ECMWF? In the second case, do PET forecasts use the same method as the one used for the historical period? → For the forecasts, we used PET data directly from ECMWF, which was computed using surface energy balance. The Penman-Monteith (PM) method requires several weather variables (such as vapor pressure, solar radiation etc.) to compute PET. However, some of these variables are not available as seasonal forecasts and therefore it was not possible to recompute the PET forecasts using the PM method.

L176: "45 ensemble members (…) were also selected from (…)" since, in my understanding, there is no generation involved. → We have corrected this sentence. (Page 5. Line 182-183)

L177-180: Here, you first mention the construction of the ESP where each member is simulated, and then mention the parameter estimation of the hydrological model. It might be more intuitive to mention the parameter estimation before mentioning simulations. → We moved those sentences as you suggested. (Page 5. Line 179-185)

L182-184: This sentence shows the issue I have with the “skill” terminology. “The Continuous Ranked Probability Skill (CRPS) method”: the CRPS is a score and not a skill, it stands for Continuous Ranked Probability Score. “method” may probably be removed. → We used the term ‘score’ when discussing CRPS, and use ‘skill’ when referring to CRPSS.

Figure 2: Here the issue with the “skill” appears clearly. Skill should come from the comparison of CRPS values corresponding to two different systems. Here instead, a skill is linked to a single CRPS box, which does not make sense given the definition of skill. In addition, the method used to calculate PET could appear to clarify the point mentioned above. → We have modified the general terminology across the revised manuscript as well as Figure 2. (Page 5. Line 189)

L222: “a water balance module” and “the United States” → We have corrected them. (Page 6. Line 237)

L226: “see Table S1” → We have corrected it. (Page 7. Line 241)

L234-235 “higher performance”: what is meant by “performance” here? Each objective function will provide good model performances as long as we focus on the flow characteristics that the objective function focuses on. → To clarify this, we have modified the sentence. (Page 7. Line 249)

L244-247 NSE formulation: the NSE usually compares the simulation to the average of observations and not to the average of simulations. → We have corrected the typo. (Page 7. Line 259-262)

L260 “the entire range of the parameter of interest”: what do you mean by this? What is the parameter of interest? Do you mean “forecast range”? Please clarify. → We have revised the sentence as you suggested. (Page 7. Line 275)

L268: This goes against the definition of the skill and of the CRPS. The CRPS alone does not provide an estimate of the skill. → The terminology issue has been corrected across the manuscript.

L271-272 “the quality of the skill”: This phrase does not make sense to me. “Quality” is what would be conveyed by the CRPS while “skill” is what is conveyed by the CRPSS. The skill is a ratio of quality/performance indices→ We have removed this sentence.

L275: “The major reasons” → We have removed this expression.

L283-286 “is more skilful than the benchmark”: A forecast system alone can only have skill with respect to a benchmark. Therefore, we can either say “the system gives higher performances than the benchmark” or “the system has skill with respect to …” (the two being equivalent). Similarly, the forecasting system and the benchmark cannot have the same skill. Lastly, a score of 1 does not necessarily guarantee a perfect forecast, if the benchmark is of sufficiently poor quality. → We agree with your point and have changed the wording as you suggested across the revised manuscript.

L287: Usually it is not the CRPSS that is averaged due to the reasons mentioned by the authors. Rather the CRPS that is averaged over all years, and the CRPSS that is computed based on the two averaged values. → In this study, we use a metric that we introduced in a previous study (Lee et al 2023), named 'overall skill', which measures the frequency which SFFs outperform ESP. So, the overall skill is not an averaged CRPSS but a probability that SFFs have skill with respect to the benchmark over the entire period and catchments.

L290 "more skillful than the benchmark": please rephrase → We have rephrased these sentences across the revised manuscript considering your previous comments (L283-286).

L293 "more skillful than ESP": please rephrase. → We have rephrased these sentences across the revised manuscript considering your previous comments (L283-286).

L305: Have you identified a reason for this gap in the last three catchments? Is there a distinctive non-stationary behavior in these catchments? Or are the processes particularly hard to model with the Tank model? → We could not find exact reason for the gap for those three catchments. However, we think it is related to the characteristics of those catchments (Imha: the driest, Namgang: the wettest, Boryung: the smallest catchment size). We have provided additional information on their characteristics. (Page 9. Line 320-323)

L310 "theoretical skill measured by the mean CRPS": please rephrase. → We have changed the terminology across the manuscript.

L318: It would be interesting to know why this catchment stands out. → Thank you for this comment. Imha is the driest catchment among all 12 catchments with the lowest modelling performance. Additional explanations are included in the sentence. (Page 9. Line 334-336)

Section 3.2: The results shown in Figure 5 are valuable and could help interpret the results of the comparison between SFFs and ESP if it was shown for bias adjusted variables. Figure 6 proves that the sensitivity of the skill to weather forcings is distorted due to biases. Why not show the bias adjustment first and then only the sensitivity to weather forcings so that this analysis can more easily feed the rest of the article? → Figure 5 shows the contribution of each weather variable to the performance of SSFs based on CRPS (i.e., there is no comparison to ESP as seen in the modified Figure 2). In addition, we aimed to demonstrate how the contribution of each variable to the performance of SFFs changes with the application of bias correction (before and after simultaneously). Therefore, we provided the results without bias correction in the manuscript and included the bias-corrected results in the supplementary material (Figure S7).

Figure 5: There is something I do not understand in the results in Figure 5. Assuming that the relative skill represented corresponds to the overall skill resented in the Methodology section, and that the benchmark in the CRPSS is the SFF with all uncertainties (forecasts of P, T and PET). Given that precipitations are key features, replacing forecast precipitation with the observed precipitation (in skill-T and skill-PET of Figure S2) should increase the performance with respect to the benchmark (greater CRPS than that of the benchmark), and should therefore give CRPSS values greater than 0 and an overall skill greater than 50%. Here, the inverse is observed. Could you please clarify this?

→ This misunderstanding is caused by the terminology. Since Figure 5 shows the contribution (%) of each weather variable to the performance of SFFs computed using CRPS, so there is no comparison with a benchmark. Thus, the increase or decrease in the area of each shape does not necessarily indicate an increase or decrease in performance (i.e., it represents the contribution rate (%) of each variable to the performance of SFFs). To make this clear, we modified Figure S2 in supplementary material.

L335-337: a word may be missing. → We have revised that sentence. (Page 9. Line 347-350)

L345: "which in reality" → We have corrected the sentence. (Page 9. Line 358)

L348-352: It appears clearly that the skill is degraded in some catchments, for some lead times, which could be fine if the average over years were to increase. However, this is not systematic based on Figure S3. Could the authors please comment on this and maybe add a point in the discussion section or in the Methodology section (L199-201)? → For most catchments, bias correction of weather forecasts enhances the overall skill. As you mentioned, in some catchments and lead times, the overall skill slightly deteriorates after correcting biases. We have added a figure supports this in the supplementary material (Figure S6) as well as additional explanations in the manuscript. (Page 10. Line 366-371)

L371-373: The range between 45% and 55% is somewhat subjective. I would recommend applying statistical tests instead to ensure that the full distributions of CRPS are statistically different, for instance. → In this study, we used the concept of 'overall skill' representing the probability that SFFs outperform ESP for certain period of time for a given catchment. We believe the overall skill might provide the performance of SFFs more intuitively. The range that we used here (±5%) seems to be somewhat subjective, however, even with the statistical tests, there may still be a need for subjective choice regarding the level of confidence. We have provided additional explanations for the reason in the modified manuscript. (Page 11. Line 389-393)

L403 "average years": Isn't Figure 8a showing results for all years, and not only average years? → We have corrected it. (Page 12. Line 428)

L408: I suggest "all years" instead of "entire years" → We have corrected it as you advised. (Page 12. Line 417)

Figure 8: Could you please indicate the number of points that is shown behind the lines? Is it the number of catchments? The number of catchments x the number of years x the number of months in the season? Over how many points is the overall skill computed? Is it statistically representative? Would it be possible to show catchments that stand out and relate it to the analysis in Section 3.3?

→ Here, the pale black points represent the overall skill for all seasons for each catchment and this is shown in the legend below the figure. The overall skill averaged over 2011-2020 (Figure 8a) for all seasons (black line) is computed using 10,080 data (12 catchments x 12 months x 10 years x 7 lead times). In addition, Figure S8 in the supplementary material shows the detailed results (as overall skill rank) for each catchment and it can be related to the analysis described in Section 3.3 of the manuscript.

Section 3.5: It would help the reader to know the CRPS or skill obtained in this catchment. In addition, an underestimation in wet years and an overestimation in dry years are observed. Could the authors comment on this? → We have added explanations describing the overall skill of the Chungju catchment and the features of underestimation and overestimation. (Page 12. Line 435-437, Page 13. Line 445-446)

L424: On Figure 9, it seems obvious for the 1-month cumulative flows, but not necessarily for the other time periods. → We have revised the sentence to clarify this. (Page 13. Line 448-449)

L462: This point is very relevant. It would be interesting for readers that are not familiar with the area whether ENSO is a good predictor in South Korea. → To enhance this point, we have added explanations on the significance of ENSO in the skill of seasonal forecasts, along with insights into the connection between regional weather patterns and ENSO in South Korea, as supported by previous studies. (Page 14. Line 488-493)

L496: "useful; however" → We have corrected it. (Page 15. Line 527)

L507-508 "we investigated the skill of seasonal weather forecasts": in my understanding, all analyses focus on the skill of seasonal flow forecasts. → We have modified this sentence to convey our message more clearly. (Page 15. Line 540-541)

L511: "have not been tested" → We have corrected it. (Page 15. Line 544)

L511 "more broadly research": I am no sure this is correct. Please consider rephrasing. → We have modified this sentence. (Page 15. Line 550)