REPLY TO REFEREE#2 COMMENTS

This contribution by Lee et al. presents a performance assessment of seasonal flow forecasts generated using ECMWF SEAS5 forecasts and the Tank hydrological model, upstream 12 operational reservoirs in South Korea.
After introducing the experimental setup, the data, the hydrological model and the evaluation framework, the authors analyse the skill of seasonal flow forecasts. First the authors assess the sensitivity of the skill to the hydrological model performance, to the model inputs, namely P, PET and T, and bias correct these inputs. Then they assess the skill of seasonal flow forecasts generated based on SEAS5 with respect to the standard ESP method, distinguishing dry and wet seasons as well as dry and wet years. Lastly the authors show an example of flow forecasts in a given catchment.
Overall, this paper is well structured and written, though several typos remain in some parts of the text, which I tried to list below. The figures are relevant, informative and well presented, though the captions could sometimes be more detailed. The methodology and the analysis of results are both comprehensive and methodical. However, I list hereafter three concerns, the major one being the definition of skill in the manuscript. These are followed by a list of minor comments, mostly asking for clarifications and some reformulating.
Based on this, I recommend the paper for publication subject to major revisions as this work will be a valuable insight into the application of seasonal forecasts over South Korea with an extra focus on reservoir management.

We thank you for taking the time to review our manuscript and are grateful for the positive comments. We are committed to address them in our revision.

General comments

1. The term “skill” in the article is used with different meanings, and sometimes with meanings that are not consistent with the definition commonly used in the literature (see e.g. the books by Wilks, or Jolliffe and Stephenson). The skill in essence is the comparison of the performance of a forecast system with the performance of a benchmark (e.g. ESP is used here) as in Eq. 9 of the manuscript, which I refer to as “skill” in this review. This ratio ranges between -infinity and 1 (see the books by Wilks and Jolliffe and Stephenson for instance). The authors here re-employ the “overall skill” from a previous paper as the percentage of years during which the forecast system has skill with respect to the benchmark, introducing a sense of variability which is interesting. However, in the results section this becomes confusing as the authors use in turn: “actual skill” and “theoretical skill” both having values greater than 1 or 100%, “skill ratio” which is smaller than 1, “relative skill” which ranges between 0 and 100%, “overall skill” which is clearly defined, yet the term is somewhat misleading. It seems important to clarify this aspect for the paper to be understandable, to ensure scientifically sound conclusions. A non-exhaustive list of instances where this was unclear is provided in the detailed list hereafter.

   We appreciate your comments. A similar point was also raised by Referee 1. We agree that our terminology was unclear and will clarify this important point in the revised manuscript, using ‘score’ when referring to CRPS and ‘skill’ when referring to CRPSS (which uses ESP as benchmark).

2. Related to this first point, the skill thus allows the comparison of two forecasting systems. In section 3.1 it would thus seem natural to look at a skill where the numerator is the CRPS computed against pseudo-observations, and the denominator is the CRPS against real observations (or vice-versa). The result should range between -infinity and 1 if the authors

use the skill, or between 0 and 100% if they use the "overall skill". The same reasoning applies to the comparison before and after bias correction, to the experiment of skill from weather forcings, and to the comparison with the ESP (see for instance the methodologies of Crochemore et al. 2020 and Greuell et al. 2019). However, it was not entirely clear if this is what was systematically done, and I suggest clarifying this for each of the results section and in the figure captions.

We will revise the terminology across the manuscript. In addition, to clarify this issue, we have modified the schematic diagram Figure 2 (see page 6 in this document).

3. The authors did not exploit much the spatial heterogeneity in catchments, though they do mention that no correlation could be found in terms of skill with respect to the ESP (Section 3.3). I still wonder if explanations could be given regarding the Imha, Buan, and Namgang catchments which stand out in Sections 3.1 and 3.2. The Chungju catchment is also later used to illustrate forecasts. It would be useful to understand the variability that can be found between these catchments to explain differences found in the analysis. Here as well, detailed comments and suggestions are provided hereafter.

   Thank you for this comment. We will incorporate further spatial characteristics of those catchments in the manuscript (Section 3.1 and 3.2). Additionally, we will include the skill of the Chungju catchments in Section 3.5. We believe that this modification will help readers with a more comprehensive understanding.

Wilks, D. S., 2006: Statistical methods in the atmospheric sciences. Academic Press,.
Jolliffe, I. T., and D. B. Stephenson, 2003: Forecast Verification: A Practitioner's Guide in Atmospheric Science. John Wiley & Sons Ltd., 240 pp.
Crochemore, L., M.-H. Ramos, and I. G. Pechlivanidis, 2020: Can Continental Models Convey Useful Seasonal Hydrologic Information at the Catchment Scale? Water Resources Research, 56, e2019WR025700.
Greuell, W., W. H. P. Franssen, and R. W. A. Hutjes, 2019: Seasonal streamflow forecasts for Europe – Part 2: Sources of skill. Hydrology and Earth System Sciences, 23, 371–391.

Specific comments
Abstract: There are some typos in the abstract. I suggest some modifications below but invite the authors to screen the text for typo correction. → We thank you for this comment. Once again, we will check the typos and correct them.

L14: Replace "to link" with "to generate" → We will replace it.

L15-16 "at finer scales such as catchment": A word seems to be missing → We will correct it.

L16: "generating SFFs (…) remains challenging" → We will correct it.

L19: "at catchment scale" → We will correct it.

L20: "over the last decade" → We will correct it.

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

→ We agree with you. To clarify our methodology and goal, we will replace the term 'actual skill' with 'overall skill' in the abstract and add explanations on the 'overall skill'. We will also be consistent with the terminology when using skill or score.

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 will clarify this sentence.

L30-31: are these "openly available"? → Yes, it is. We will change the term 'freely' to 'openly'.

L57: this is the first occurrence of ESP, please explicit the term. → We will add full explanation.

L58: "by forcing a hydrological model with historical meteorological observations" → We will correct this.

L67-68: "Some of these studies focused" → We will correct it.

L95: "did not analyse" → We will correct it.

L108: "may be considered" → We will correct it.

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 agree with your comment and will correct it.

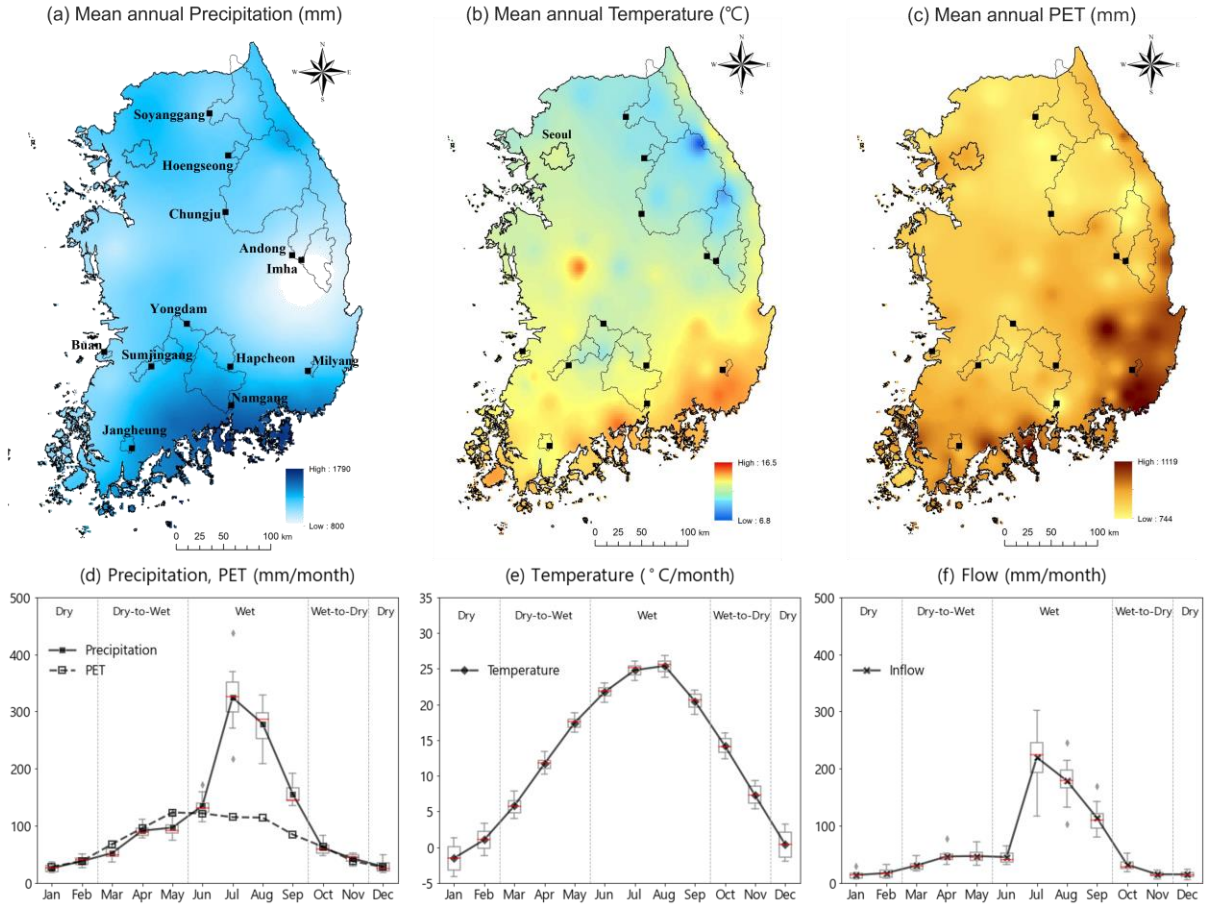
Table 1: It would be informative to add Tmin and Tmax to this table, especially given that you have catchments with snow which you later discuss. → We agree with your comment and have added Tmin and Tmax in Table 1 – see modified version below.

Modified Table 1. Mean annual (2001-2020) properties of the 12 multipurpose reservoirs (from north to south) and the catchments they drain (K-water, 2022). Here, Tmin and Tmax represent mean monthly minimum and maximum temperature averaged over 2001-2020. (P: precipitation, T: temperature, PET: potential evapotranspiration)

Catchment		Soyanggang	Hoengseong	Chungju	Andong	Imha	Yongdam	Buan	Sumjingang	Hapcheon	Milyang	Namgang	Jangheung
Area (km^2)		2703	209	6648	1584	1361	930	59	763	925	95	2285	193
Mean annual	P (mm)	1220	1336	1197	1079	956	1317	1292	1343	1279	1375	1477	1439
	T (℃)	10.8	10.9	11.1	11.1	12.2	11.8	13.5	12.6	12.8	14.2	13.5	13.8
	T min	-4.2 (Jan.)	-4.0 (Jan.)	-3.2 (Jan.)	-3.5 (Jan.)	-1.6 (Jan.)	-2.3 (Jan.)	-0.1 (Jan.)	-1.5 (Jan.)	-0.8 (Jan.)	1.0 (Jan.)	0.4 (Jan.)	1.3 (Jan.)
	T max	24.0 (Aug.)	24.1 (Aug.)	25.9 (Aug.)	23.8 (Aug.)	25.1 (Aug.)	24.8 (Aug.)	26.7 (Aug.)	25.8 (Aug.)	25.5 (Aug.)	26.8 (Aug.)	26.0 (Aug.)	26.2 (Aug.)
	PET (mm)	874	870	881	896	947	884	960	919	933	993	952	896

(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. → Thank you for this suggestion. To show the inter-catchment variability in the figure, we have

added box plots in Figure 1 (d, e, f) shown below.



Modified Figure 1: Top row: mean annual (1967-2020) (a) precipitation (mm/year), (b) temperature (℃/year) and (c) PET (mm/year) across South Korea and the boundaries of the 12 reservoir catchments analysed in this study (all maps obtained by interpolating point measurements using the inverse distance weighting method). Bottom row: (d) cumulative monthly precipitation and PET, (e) mean monthly temperature and (f) cumulative monthly flow. All variables are averaged over the 12 reservoir catchments from 2001 to 2020. Box plots show the inter-catchment variability.

In addition, in the caption: L133 "Mean monthly": isn't it rather the sum in the case of precipitation, PET and flow? → Agree. We have revised this error. See revised caption above.

L135 "variability of each weather variable": "of each weather and hydrological variable". Is it the inter-catchment variability or the inter-annual variability? → We have modified Figure 1 as shown above. Now, it represents the inter-catchment variability, and we will clarify this in the revised manuscript.

L149: Please introduce the abbreviation KMA here. → We will add full explanation of KMA.

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? → Thanks to your comment, we realised that this sentence might be unclear and could lead to misunderstandings. In this study, when we refer to 'flow data,' we specifically mean the flow to the reservoir from their upstream catchment, estimated by K-water. We meant to say that 'K-water generates flow data using the water balance equation; however, reservoir evaporation is not considered in this process.' We will change the term 'streamflow' to 'flow' and clarify this point in the manuscript.

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. → Agreed. We will add the reference in that sentence.

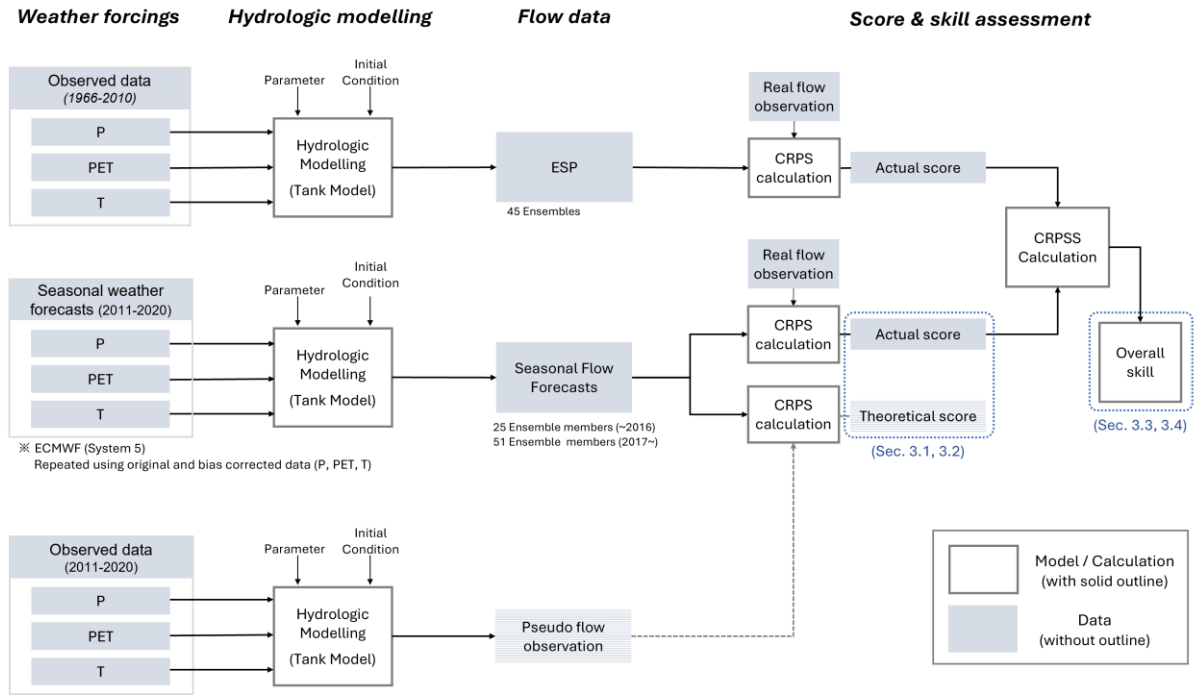
L174: Did you compute PET based on the Penman-Monteith method as mentioned L151, or did you retrieve PET forecasts directly from ECMWF? In the second case, do PET forecasts use the same method as the one used for the historical period? → Thank you for this comment. For the forecasts, we used PET data 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 agree with your comment and will correct this sentence accordingly.

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. → Thank you for this suggestion. We will reorder those sentences as you suggested.

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 agree with your comment. We will use '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. → Thank you for this comment. We have modified the terminology and Figure 2 shown below. As we replied above, we used PET forecasts provided by ECMWF.



Modified Figure 2: Schematic diagram illustrating analysis method of the study.

L222: "a water balance module" and "the United States" → We will correct it.

L226: "see Table S1" → We will correct it.

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. → Thank you for this comment. We intended to convey that among many possible objective functions, this objective function (Eq.4 in the manuscript) showed the best results in calibrating the Tank model as suggested by a previous study (Kang et al., 2004). To clarify this, we will make it clearer in the manuscript.

L244-247 NSE formulation: the NSE usually compares the simulation to the average of observations and not to the average of simulations. → Thank you for this correction. We will correct the typo.

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. → Yes, we mean "forecast range". We will revise the sentence as you suggested.

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 will be 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→ Thank you for this comment. We will modify this sentence.

L275: “The major reasons” → We will correct it.

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 will change the wording as you suggested.

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. → Thank you for your comment. 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. We will modify this sentence in the manuscript to make it clearer.

L290 “more skillful than the benchmark”: please rephrase → We will rephrase this sentence considering your previous comments (L283-286).

L293 “more skillful than ESP”: please rephrase. → We will rephrase this sentence 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? → Thank you for this comment. 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 will provide additional information on their characteristics.

L310 “theoretical skill measured by the mean CRPS”: please rephrase. → We will change the terminology across the manuscript.

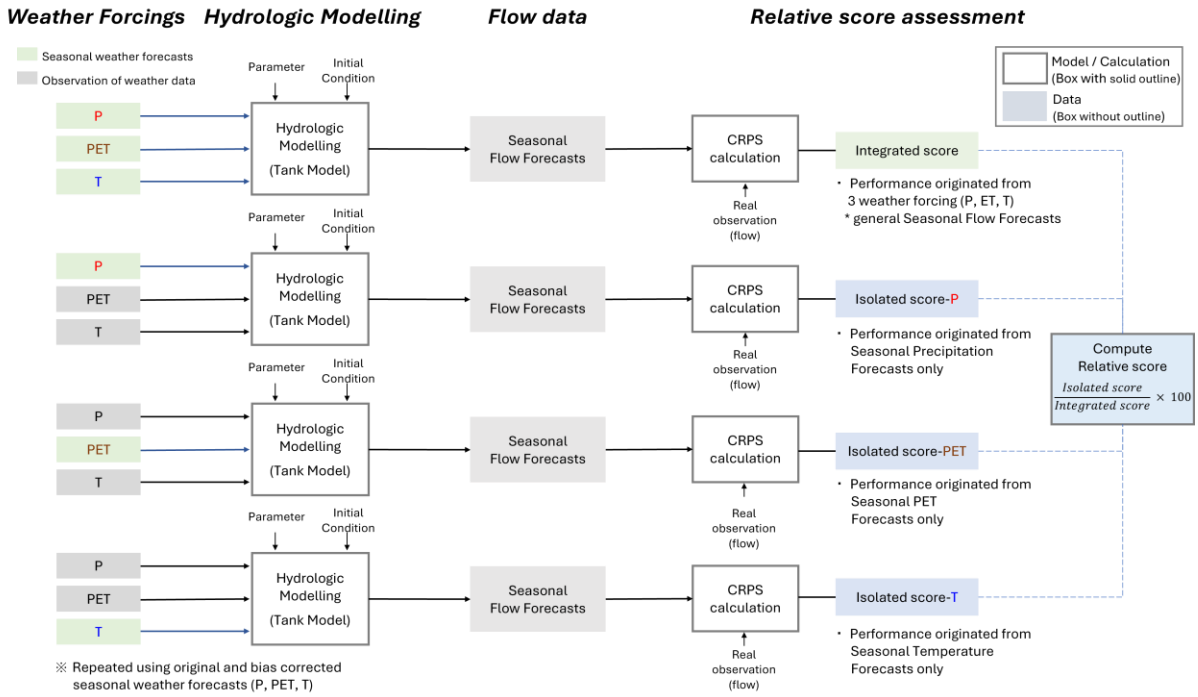
L318: It would be interesting to know why this catchment stands out. → Thank you for this comment. Imha is the driest catchment among all 12 catchments with the lowest modelling performance. Additional explanations will be included in the sentence.

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? → Thank you for this comment. Firstly, 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 S6).

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?

→ We thank you for this comment. 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 as presented below.



Modified Figure S2: Schematic diagram of calculating the relative performance.

L335-337: a word may be missing. → We will revise that sentence.

L345: "which in reality" → We will correct that sentence.

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)? → We thank you for this suggestion. 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. It is clearly shown in the figure below (Figure R1), and we will add this figure in the supplementary material as well as additional explanations in the manuscript.

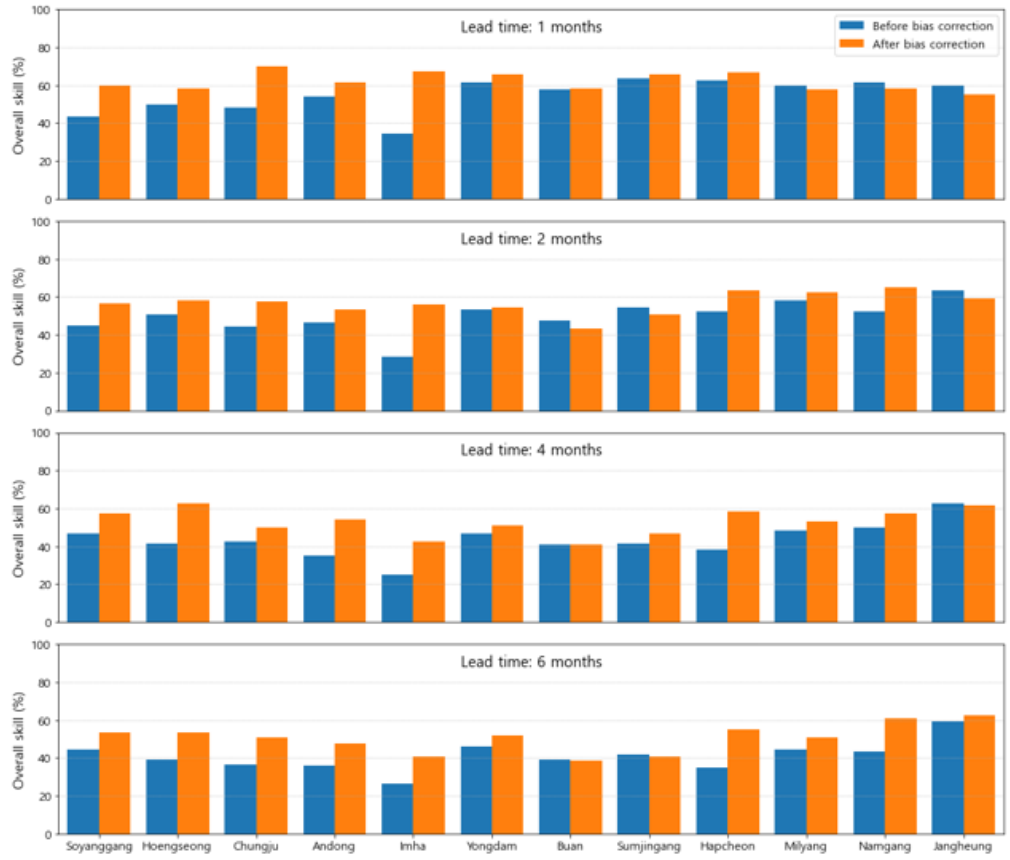


Figure R1: Overall skill comparison for each catchment before (blue bar) and after bias correction (orange bar) of weather forcings (P, T and PET) at lead times of 1, 2, 4, 6 months (from top to bottom).

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. → Thank you for this comment. 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.

L403 "average years": Isn't Figure 8a showing results for all years, and not only average years? → We will correct it.

L408: I suggest "all years" instead of "entire years" → We will correct it.

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?

→ We thank you for this comment. 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 S6 in the supplementary material (shown below) 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.

		(a) Average over 2011 – 2020							(b) Average over dry years (2015, 2017)							(c) Average over wet years (2011, 2020)						
	Lead time	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
North	Soyanggang	7	7	6	4	4	4	7	11	8	10	9	9	11	10	8	11	6	4	4	3	2
	Hoengseong	8	5	5	1	6	4	3	5	3	5	6	6	7	5	2	1	4	1	4	6	6
	Chungju	1	6	9	9	7	7	6	1	8	12	11	9	7	6	2	5	2	4	2	3	2
	Andong	6	10	8	6	8	9	9	5	6	2	2	1	3	3	11	5	1	3	7	8	10
	Imha	2	8	11	11	11	10	11	5	7	10	9	11	7	9	5	5	8	9	8	7	6
	Yongdam	4	9	6	8	9	6	8	3	8	7	8	8	6	6	10	11	8	7	10	8	10
	Buan	8	12	12	12	12	12	12	3	12	8	11	12	12	12	11	10	12	12	10	11	8
	Sumjingang	4	11	10	10	10	10	9	5	11	8	7	6	10	10	1	9	11	11	10	12	10
	Hapcheon	3	2	3	3	3	3	4	1	3	2	1	1	1	1	5	3	8	9	8	8	9
	Milyang	11	3	4	7	4	7	5	9	2	2	2	1	4	6	2	1	2	6	4	3	2
	Namgang	8	1	1	4	2	2	2	10	1	1	2	4	2	2	8	3	4	7	3	1	2
South	Jangheung	12	4	1	2	1	1	1	12	3	6	5	5	4	3	5	8	6	2	1	1	1

Figure S6: Overall skill ranks for each catchment averaged over (a) entire years (2011 to 2020), b) dry years (2015, 2017) and (c) wet years (2011, 2020) for all seasons (January to December). The catchments are arranged from the top to bottom in order of their location from the northernmost (Soyanggang) to the southernmost (Jangheung). The three most (least) skilful reservoirs are highlighted in yellow (pink) colour.

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? → Thank you for this suggestion. We will add explanations describing the overall skill of the Chungju catchment and the features of underestimation and overestimation in Section 3.5.

L424: On Figure 9, it seems obvious for the 1-month cumulative flows, but not necessarily for the other time periods. → We agree with you and will revise the sentence to clarify this.

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. → Thank you for this comment. To enhance this point, we will add 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.

L496: “useful; however” → We will correct it.

L507-508 “we investigated the skill of seasonal weather forecasts”: in my understanding, all analyses focus on the skill of seasonal flow forecasts. → We agree with you and will modify this sentence to convey our message more clearly.

L511: “have not been tested” → We will correct it.

L511 “more broadly research”: I am no sure this is correct. Please consider rephrasing. → We will modify this sentence (to ‘broader research’).