Dear Prof. Jinkyu Hong,

Many thanks for handling the review process for our manuscript. The time and effort devoted to our manuscript by you and the reviewers are very much appreciated.

We have revised the manuscript carefully according to the reviewers' comments and suggestions. In the following, we provide a point-by-point response. The original reviewer comments are in black regular font. Our responses are shown in blue italic font. Quotes from the revised paper are shown in blue bold-face font.

REVIEWER COMMENTS

The central question for this manuscript is whether one can robustly claim that the implementation of a multi-layer snow scheme has a substantial impact on seasonal forecasts in GloSea6. In this revision, the authors have added an additional experiment (G6single) in which the snow scheme is kept single-layer within the GloSea6 framework, so that other model updates from GloSea5 to GloSea6 are fixed except for the snow parametrization. Below I summarise my main concerns.

- ➔ *We thank the reviewer for your thorough and constructive comments, which have helped us to clarify the scope and strengthen the manuscript. Below we respond point-by-point and describe the corresponding revisions.*

Major comments

1. With the added G6single experiment, the natural way to quantify the impact of the multi-layer snow scheme on seasonal forecasts is to focus on the model differences that isolate the snow parametrization.

G6multi – G6single in the coupled system (Fig. 2) and JULESmulti – JULESsingle in the offline LSM experiments (Fig. 1).

In Fig. 1, the JULESmulti – JULESsingle differences in surface properties appear visually quite large. However, this is at least partly due to the very narrow vertical range used on the y-axes. For example, the differences in latent heat flux are generally smaller than about 1.2 W m^{-2}. It is not clear whether such small differences are statistically significant and/or physically meaningful for seasonal forecasting, especially given that typical summer LH values are several tens of W m^{-2}.

A similar issue arises for the coupled simulations in Fig. 2. The snow-scheme difference G6multi – G6singleis consistently smaller than the total difference G6multi – G5single for essentially all variables, including precipitation. This implies that a substantial part of the difference between GloSea5 and GloSea6 including meteorological conditions must arise from other updates (e.g. atmospheric, ocean, sea-ice, land-cover/albedo changes, stochastic physics and ensemble size), not from the snow-scheme change alone.

The same pattern appears in other figures. For example, in Fig. 5, visual differences in air temperature and precipitation between G6multi and G6single seem small. Accordingly, it is not clear whether they are statistically or practically significant.

- ➔ *We agree that the natural framework to isolate the effect of the snow parametrization is to analyse the pairs JULESmulti–JULESsingle and G6multi–G6single. We have revised Figs. 1 and 2 so that, for each variable, the y-axis ranges are adjusted to easily find the difference between the offline (JULES) and coupled (GloSea) experiments. This prevents the JULESmulti–*

JULESsingle differences of land heat fluxes (e.g., latent heat flux) from appearing exaggerated relative to G6multi–G6single. Furthermore, to confirm the significance of the difference when applying the multi-layer snowpack scheme in the offline and coupled experiments, a significance test is conducted, and the results are explicitly marked in the time series in Fig. 1 and Fig. 2. The description of statical significance in the time series of climatological differences in Lines 288-293:

"To identify climatological differences between single- and multi-layer snowpack schemes in offline and coupled experiments, statistical significance is tested using all samples (i.e., all years and ensembles) with the Student's t-test. The statistical significance in the time series of the differences (Figs. 1 and 2) is assessed within a ±5-day window centered on each calendar date, and a False Discovery Rate (FDR) corrected t-test (Benjamini–Hochberg) is used at the 10% level across the spatial grid to prevent the inflation of false positives, thereby ensuring the statistical robustness in the spatial domain of the differences found (Figs. 1, 3, 5, and 7)."

➔ *For the offline JULES experiments, we apply a two-sided Student's t-test to the area averaged difference between JULESsingle and JULESmulti over the 22-year period (2001-2022) within an 11-day window centered on the calendar date. For the coupled simulations, we perform analogous tests on the variables simulated by 24-year hindcast ensemble runs initiated on 1 March. These results show that, although the absolute values of latent heat flux differences of about 1 W/m² are small compared with peak summer values, due to their persistence they are statistically robust over large, coherent regions in the snow-frontal zones. We explicitly state that process-based effects, not just the peak magnitude of flux differences at a given day, in Lines 438-447.*

"While the additional 1 W/m2 of latent heat flux appears marginal, it is critical to consider the accumulated effect over the seasonal forecast period. A small anomaly can be significant when persistent, in the context of land-atmosphere coupling. For instance, a persistent difference of 1 W/m2 in latent heat flux over one month translates to a cumulative change of ~1 mm in the water budget. Such an alteration in the regional water and energy budget is physically meaningful and can serve as a non-negligible source of memory and predictability in precipitation. To illustrate the physical sequence between land surface variables by the realization of snow physics, the lead-lag correlation of major water budget variables is compared between G6single and G6multi (Fig. 2j). The results show the hydrological chain of SSM→LH→PR with a positive correlation among variables in each segment, characterized by a lead-lag time of approximately one week. In other words, the increased soil moisture in mid-latitude regions likely increases precipitation based on positive evapotranspiration-precipitation feedback."

2. When comparing Fig. 1 (offline JULES) and Fig. 2 (coupled GloSea) the snow-scheme differences clearly show changes in both magnitude and timing (e.g. the peak period of the anomalies). In particular, the snow-scheme differences in the coupled system (i.e., G6multi – G6single) appear smaller than those in the offline LSM, and their peaks are shifted.

Taken together with the fact that G6multi – G6single is much smaller than G6multi – G5single, these results suggest that the majority of the differences between GloSea5 and GloSea6 are likely due to updates in ensemble size and atmospheric/ocean/sea-ice physics and other components, rather than to the snow-scheme change alone.

➔ *As mentioned above, to confirm the significance of the differences between the offline and coupled experiments, the result of statistical significance is included in Figs. 1 and 2. As a result, in offline simulations, the multi-layer snow scheme contributes to statistically significant changes only in the snow-covered season, while in coupled simulations, significant changes are observed not only in the snow-covered season but extend into the subsequent summer season. Because the magnitude of the changes in land surface variables is relatively small by implementing multi-layer snowpack scheme, it was difficult to understand whether the changes are significant. So, information about the statistical difference is added in this revision. However, because the differences due to phase shifts in GloSea5 and GloSea6 are larger, the impact of other updates was greater in GloSea6, which is now noted in Lines 450-454:*

"The difference between G5single and G6multi consistently exceeds the isolated snowpack scheme difference across most variables. The substantial difference between G5single and G6single confirms that updates other than the snow scheme contribute significantly to the climatological mean change in the simulation of land surface variables. However, the core finding of this study is the demonstration that the implementation of the multi-layer snow scheme yields a statistically significant and physically consistent impact that is independent of these other updates."

3. The critical p value for rejecting the null hypothesis in the Granger-causality analysis is 1−p > 0.5, i.e. p < 0.5. This threshold is much weaker than typical statistical analysis (e.g. p < 0.1 or p < 0.05) and does not provide strong evidence for causality.

➔ *Regarding the threshold value to determine rejecting the null hypothesis in the Granger-causality, 0.5 was written incorrectly in the previous version: it should have been 0.05. In the revised manuscript, it is corrected. We now interpret the quantity 1–p as a continuous measure of predictive precedence; in the text we describe only regions where 1–p is relatively large as areas with strong evidence of Granger causality.*

4. The title of the manuscript explicitly refers to a "seasonal forecast system", which naturally leads the reader to expect a quantification of seasonal forecast skill(e.g. correlation, RMSE, probabilistic scores) for the different model configurations. In the current version, however, the focus is primarily on mean-state differences and process diagnostics, and the quantitative assessment of seasonal forecast skill remains limited.

➔ *We agree that the original wording of the title could create the expectation of a detailed forecast skill evaluation. Our primary objective is to evaluate how the multi-layer snow scheme affects climatological biases and the fidelity of land–atmosphere coupling processes in a model that is used as an operational seasonal forecast system, rather than to document forecast skill in detail. To better reflect this, we have revised the title to* ***"Implementation of a multi-layer snow scheme in a seasonal forecast system: Impacts on land–atmosphere interactions and climatological biases"****. We hope this makes clear that (i) the system is used for seasonal prediction, and (ii) the focus is on coupling and climatology rather than on skill metrics. In the section of Summary and Conclusions, we explicitly note that improvements in mean-state climatology and land-surface processes do not necessarily translate into large improvements in forecast skill.*

5. In several figures, gridpoint-wise significance tests are applied to the differences between model configurations (e.g. t-tests for mean differences, Granger causality p-values) and the results are displayed on spatial maps. Even though these comparisons are model–model rather than model–observation, this still constitutes a multiple-testing problem, because many hypothesis tests are performed simultaneously across the spatial grid.

In such a setting, a certain fraction of grid points will appear "significant" purely by chance even if there is no true signal.

- ➔ *The reviewer raises a valid point regarding the application of multiple hypothesis tests across the spatial grid, which can lead to a certain fraction of falsely significant grid points purely by chance (the False Discovery Rate problem). Aforementioned in response to the first reviewer's comment, to address this crucial concern and ensure the robustness of our findings, we have applied a stricter statistical procedure to our spatial significance maps: the False Discovery Rate (FDR) procedure (Benjamini–Hochberg) at the 10% level across the entire spatial domain. The application of the FDR procedure resulted in a more stringent criterion for statistical significance, particularly reducing the number of isolated significant grid points. However, we found that the spatial patterns of significance, especially over the key snow-frontal regions and mid-latitudes, remain consistent with our original findings. Only grid points that remain significant after FDR control are stippled in the revised figures (Figs. 1, 3, 5, and 7). We have replaced phrases such as "significant at the 95% level" in contexts where only raw gridpoint tests were previously used, by specifying that significance is "at the 95% level after FDR control across the grid".*

6. Before the G6single experiment was introduced, the correlation and causality diagnostics (e.g. R(SSM,LH), R(Rn,LH), Granger causality maps) were arguably the only way to infer the possible role of the snow scheme. With G6single now available, the primary evidence for the snow-scheme impact should come from the direct model differences (G6multi – G6single and JULESmulti – JULESsingle).

At present, it is not fully clear how much additional support the correlation and causality diagnostics provide for the claim that the multi-layer snow scheme has a substantial impact in GloSea6, beyond what can be inferred from the direct differences.

- ➔ *We agree that the direct differences serve as the primary and necessary evidence for attributing changes in the coupled system to the multi-layer snow scheme. However, the diagnostics are included because they provide essential additional support by offering mechanistic validation and assessing the model fidelity of the simulated land-atmosphere coupling processes.*

- ➔ *The direct difference plots (e.g., Fig. 3f) demonstrate that the snow scheme causes wetter soil moisture and reduced temperature bias. The diagnostics, however, help reveal the physical mechanisms by which this change is achieved. The correlation metrics (R(SSM,LH) and R(Rn,LH) prove that the transition to a wetter state in G6multi leads to a fundamental shift in the land-atmosphere coupling regime—specifically, a weakening of the water-limited coupling and an enhancement of the energy-limited coupling. This physical closure is critical for interpreting the subsequent reduction in the near-surface warming bias through increased evaporative cooling. The result shows that G6multi not only reduces the bias in mean temperature but also achieves an improved spatial correlation and magnitude of the observed coupling features (e.g., Fig. 7g). This demonstrates an enhanced fidelity of model land-atmosphere coupling, proving that the multi-layer snow scheme improves the reliability of the underlying physical processes, which is a key requirement for reliable forecast systems. This is* *added in Lines 356-357 and 616-618:*

“While direct differences between G6multi and G6single isolate the mean state impact, these metrics provide process-based validation by assessing the model's fidelity in simulating the underlying processes.”

“This shift demonstrates a robust improvement in the underlying land-atmosphere coupling processes, leading to a better simulation of near-surface atmospheric variables (namely temperature and precipitation).”

➔ *Furthermore, the Granger causality analysis demonstrates the explicit linking the improved land surface states (wetter soil->higher evaporative fraction) to the atmospheric response (increased precipitation). This supports the claim that the snow scheme has a substantial impact by improving the simulated evapotranspiration-precipitation feedback loop, providing a physically coherent explanation for the improved precipitation distribution in G6multi (Fig. 5l).*

Minor comments

1) For clear comparison between the offline and coupled experiments, y-axis scales used in Fig. 1 and Fig. 2 for the same variables should be the same. Otherwise, small differences may appear exaggerated in one figure and muted in another.

➔ *We agree and have revised Figs. 1 and 2 so that the y-axis limits are identical between the JULES and GloSea panels. In Fig. 2, we tried to show the difference between G6multi–G6single as well as G6multi–G5single, so I couldn't make the scale identical to Fig. 1, but we adjusted the scale of the right y-axis of Fig. 2g,h so that comparison is possible.*

2) The text around line 385 refers to Fig. S2 for evidence that differences in initial conditions are negligible. However, there are no relavant information to show quantitative differences in initial conditions to drive the climate model. Please revise Fig. S2 (or add a new supplementary figure or table) to provide such quantitative evidence or adjust the text accordingly.

➔ *We would like to clarify that the raw initial condition (IC) for the G5single experiment is currently unavailable due to data archival limitations, which prevents us from directly plotting the IC differences between GloSea5 and GloSea6 in the Figure S2. However, to address the reviewer's concern and verify the assumption that IC differences are negligible, we quantitatively analyzed the 1-day forecast as a robust proxy. Since land surface variables evolve relatively slowly, the 1-day forecast effectively represents the initial state particularly in snow variables. Our analysis of the multi-year runs confirms that the differences in these fields are statistically insignificant across the domain. We have revised the corresponding text (Lines 396-397) to explicitly include this quantitative justification.*

“…an analysis of 1-day forecast fields, which serve as a robust proxy for the initial land state due to their slow evolution, confirms that the difference in initial snow amount is statistically insignificant (Fig. S2).”

3) The analysis in this manuscript is limited to the Northern Hemisphere, not the global domain. Authors may want to adjust the title to reflect this spatial focus, or to state this limitation prominently in the Abstract and Introduction.

➔ *To clarify the research domain to the Northern Hemisphere, we added the sentences to state this information in Lines 15-16 and 96-97:*

"Results show that the multi-layer configuration better reproduces the observed Northern Hemisphere snow seasonality."

"The evaluation is restricted to the Northern Hemisphere (NH) and mainly to snow-affected mid-latitude regions."

4) In Fig. 2(j), the "standardized difference" (G6multi–G6single) time series is shown, but its definition and interpretation remain unclear. It is not obvious that simply dividing the model difference by the model standard deviation provides a meaningful measure of the significance of the model differences. Please provide a precise mathematical definition (over what period and domain the standard deviation is computed) and explain what aspect of the physical behaviour this metric is intended to highlight. If the goal is to emphasise lead–lag relationships between variables, you may consider presenting or at least explicitly referring to lead–lag correlations instead.

➔ *Based on the reviewer's suggestion, we include lead–lag correlations between the differences (G6multi–G6single) in soil moisture, latent heat flux, and precipitation in Fig. 2j. The results demonstrate that positive soil moisture differences tend to precede latent heat flux differences by about one week; and latent heat flux differences tend to precede precipitation differences by about one week. We document these findings in Lines 442-447:*

"To illustrate the physical sequence between land surface variables by the realization of snow physics, the lead-lag correlation of major water budget variables is compared between G6single and G6multi (Fig. 2j). The results show the hydrological chain of SSM→LH→PR with a positive correlation among variables in each segment, characterized by a lead-lag time of approximately one week. In other words, the increased soil moisture in mid-latitude regions likely increases precipitation based on positive evapotranspiration-precipitation feedback."