Response to Anonymous Referee #2

We want to thank the reviewer for their effort and their very helpful comments on our manuscript. Please find below the detailed answers to all comments. The reviewer's comments are in blue, our answers in black. Text from the manuscript is highlighted in grey, and changes to the manuscript are shown in red on a grey background.

The manuscript has a strong layout and proceeds logically. The ensemble itself has been presented before (Tabone et al., 2024), although the prior analysis focused on the Northeast Greenland ice stream and interior thinning, rather than Greenland as a whole. This work is a logical extension of the ensemble and a useful contribution. The method of assessing the ensemble members was interesting and well described.

We thank the reviewer for their positive assessment and are pleased that they consider this work a useful contribution. Below we address each of the comments, which we believe will help to improve the manuscript.

My largest concern about the manuscript is the climate forcing. The authors find that the atmospheric corrections terms are quite important to ice sheet evolution, but I believe they represent a single value for all locations on the ice sheet. I'm surprised that only one combination of surface temperature and precipitation was used. The ensemble used Buizert et al. (2018) for the temperature and Badgeley et al. (2020) for the accumulation. The Buizert and Badgeley reconstructions have large differences in the spatial pattern of climate change. The lack of diversity in the climate forcing prevents fuller assessment of ice sheet model parameters. I recognize that this is not something easy to remedy after the ensemble is has been run, and don't suggest that the authors need to do it for this paper; but I do think the authors need to highlight this limitation much more strongly in the methods, discussion, and throughout the results. Although likely unintended, my primary takeaway from this paper is that the climate forcing likely matters more than any of the ice sheet model parameters.

We agree with the reviewer that the climate forcing plays a major role in the simulated evolution of the Greenland ice sheet (GrIS). This is supported by ice-sheet model intercomparison studies of future projections, where the uncertainty arising from the atmospheric and oceanic forcing is comparable to that arising from differences among ice-sheet models (Goelzer et al., 2020).

As explained in our response to Reviewer #1, Tabone et al. (2024) tested the sensitivity to the temperature forcing by running a second 3,000-member ensemble forced with the temperature reconstruction of Badgeley et al. (2020). That reconstruction produced a damped and delayed early-Holocene warming in northeast Greenland, which worsened the fit to the observed deglaciation timing. Buizert et al. (2018) was therefore selected for temperature. Since it does not provide precipitation, we used Badgeley et al. (2020), which, to our knowledge, is the only spatially complete precipitation reconstruction for Greenland covering the last 20 kyr. We selected its high-accumulation scenario because it best matches the rapid early-Holocene accumulation increase inferred at NGRIP.

We also note that the melt correction is not spatially homogeneous over the whole ice sheet: it varies with latitude through the parameter b (Eq. 7 in the original manuscript). This partly accounts for spatial biases in the forcing, although it cannot correct its full spatial pattern.

Nevertheless, we agree that using a single combination of forcings is a limitation of this study. Both the identification of the best-fit simulation and the inferred parameter sensitivities, including the SHAP analysis, are conditional on this choice. In the revised manuscript, we will highlight this limitation more clearly (same as for the first reviewer):

1) In the Method section, L150-151 are modified with these sentences:

[SAME MODIFICATION as for R1]

As developed in Tabone et al. (2024), the Buizert et al, (2018) reconstruction is selected as it best reproduces the observed timing of deglaciation onset in northeast GrIS. Since Buizert et al. (2018) does not provide a precipitation field, we take precipitation changes from the high-precipitation scenario of Badgeley et al., (2020), which best matches the rapid early-Holocene accumulation increase inferred from the NGRIP ice core. These changes are applied through the scaling relationship:

2) In the Method section, L163-L165 are substituted with these sentences:

[SAME MODIFICATION as for R1]

The ensemble consists of 3,000 transient simulations spanning the last 22kyr, in which the climate forcing is identical for all members, while 13 key parameters (Table 2) are perturbed using Latin hypercube sampling (Stein et al., 1987). For further details on model physics and experimental setup, see Tabone et al. (2024).

3) In the Discussion section, L563-L569 are substituted with these sentences:

[SAME MODIFICATION as for R1]

Finally, our results show that, for the forcing and parameter space considered in this study, the performance of the simulations in reproducing deglacial observables is primarily driven by the specific ice-sheet sensitivity to climate forcing. The selected forcings (Buizert et al., 2018, Badgeley et al., 2020) appear to effectively capture the reconstructed temperature and precipitation patterns characteristic of the deglaciation. Driving the ice sheet with these climatic fields serves as a robust approach to reduce the degrees of freedom within the system, effectively constraining the ice-sheet evolution to the observational record. However, this choice also conditions our conclusions. Several of the perturbed parameters, such as the melt corrections, the precipitation scaling and the ocean sensitivity, directly modulate the ice-sheet response to climate and may therefore partly compensate for biases in the prescribed forcing. Consequently, both the best-fit parameter values and the inferred parameter sensitivities, including the SHAP importance rankings, are specific to this forcing and may shift under an alternative paleoclimate reconstruction. Our approach, which fixes the climate forcing and samples the model-parameter space, is complementary to studies that instead sample the climate space to assess whether climate uncertainty alone can reconcile simulations with observations (Briner et al., 2020, Cuzzone et al., 2026).

In addition, the one-way forcing used here neglects the impact of ice-sheet evolution on the large-scale climate, and a clear next step would be to address the problem with fully coupled

simulations.

I also have a confession to make – I spilled a drink on my paper copy that I was editing on a flight, destroying my notes and specific comments on the text. I was added as a referee late to the review process and am now overdue myself, so I think it's more important to get the review back and keep the process rolling, than wait for me to have additional time to recreate my notes. I am sorry that this review is light on specific comments.

We are sorry to hear that the specific comments were lost. We are very grateful for the reviewer's time and for providing the review promptly, which has helped to keep the review process moving.

General Comments

Section 3.3.1 on the Emulator Sensitivity doesn't seem to add much to the paper. It's short, confusing, relies on "black box" functions, and seems largely detached from the rest of the paper. I don't think it is described at all in the methods sections, and it also doesn't seem like there is enough detail to reproduce the methodology. I understand you want to look at more complex multi-variate relationships, but this doesn't come through. So my vote is to just remove it to help keep what is a pretty long paper better focused and accessible.

We'll answer this question right after the next one, for better readability.

Figure 11 seems like the most important figure for the paper, but I find it confusing. First, I don't understand why there are both positive and negative values. Is this simply indicating positive or anti correlation for the direction a parameter is varied? If so, that doesn't seem very important to me and is just a result of the decision on how to formulate the parameter – instead, I think the relevant information is the absolute value of the correlation. Second, did I miss the assessment of statistical significance? Third, and here is where I reveal my ignorance and that I'm interpreting the whole thing wrong, but is a simple correlation a useful statistic. Doesn't this have a bias for parameters that have a linear relationship with the S scores? If I understand, you are essentially comparing the S values for each ensemble member by the parameter value used in that member. But what happens when the optimal parameter value is the middle, and either an increase or decrease in the parameter causes a decrease in the fit? The correlation would be bad, but there would be an important relationship between the score and the parameter that goes undetected. Would it be worth putting plots with the individual relationships in the supplement?

We thank the reviewer for these helpful comments, which have led us to improve the analysis presented in Fig. 11. We address each point below.

Sign of the correlations. Yes, the sign indicates the sense in which an increase in a parameter affects each score. We decided to include it because all parameters have a clear physical meaning, so the direction of the effect is informative for the physical interpretation of the results. For example, it shows that a higher ocean sensitivity improves the representation of the ice extent, whereas a higher calving rate degrades LGM performance. However, as explained below, this figure will be replaced as a result of the comments. The new figure indeed emphasizes the absolute value of the correlation through its square, the explained variance.

Statistical significance. We now include it in the revised figures. For each score-parameter pair, we test the significance of the best fit with an F-test, applying a Bonferroni correction (Dunn, 1961) for the 130 pairs. Non-significant fits are shown in grey in the R^2 matrix (Fig. 1), and the p-values are given in each panel of the scatter plots (Fig. 2).

Bias towards linear relationships. We agree with the reviewer that a correlation coefficient is biased towards linear relationships. We had inspected the individual parameter-score relationships before preparing the original manuscript and found them to be mostly linear, but we did not include them because they comprise 130 panels (13 parameters × 10 scores). Prompted by this comment, we have now carried out a more systematic analysis. For each parameter-score pair, we fitted linear, quadratic and cubic polynomials and selected the most appropriate one using the Bayesian Information Criterion (BIC, Neath and Cavanaugh, 2012; Schwarz, 1978):

$$BIC = n \ln\left(\frac{RSS}{n}\right) + k \ln(n),$$

where RSS is the residual sum of squares, k is the number of coefficients of the polynomial, and n is the number of ensemble members. The BIC penalises model complexity, so a higher polynomial degree is only selected when it improves the fit sufficiently.

The results are shown in Fig. 1 below, which displays the proportion of explained variance (R^2) of the best fit (linear, quadratic or cubic) for each parameter-score pair, together with the shape of the fitted curve. This analysis confirms that most relationships are well described by a linear fit: 100 of the 130 pairs select a linear model. For the remaining 30 pairs, the non-linear fit increases the explained variance by only about 1% with respect to the linear fit (median $R^2_best - R^2_lin = 0.01$). The improvement exceeds 5%in only four cases: S_{iso}-b, $S_{ice\ cover}$-b, S_H-b and S_{GRIP}-δ, with gains of 7-18%.

Although this analysis does not contradict our previous results, it provides a more complete picture of the parameter-score relationships. We will therefore:

1) Replace Fig. 11 of the original manuscript with Fig. 1 below.
2) Revise the accompanying discussion accordingly, although the main conclusions remain unchanged.

L403-L411

To assess the sensitivity of the GrIS deglaciation to the ensemble parameter values, Fig. 11 shows, for each score–parameter pair, the proportion of the variance of the score explained (R^2) by the best polynomial fit, with the degree (linear, quadratic or cubic) selected using the Bayesian Information Criterion (see the Supplementary Material). Several features can be extracted from this figure. First, no single parameter explains more than 10% of the variance of the global score S (maximum R^2 = 0.10, for c_1). The scores related to the present-day state (S_H, $S_{ice\ cover}$, S_{z_bed} and, to a lesser extent, S_u), and the isochrone-based score (S_{iso}) are all significantly influenced by the SMB correction parameters, both the latitude-dependent correction term b and the spatially constant correction term c_1 from Eq. 6. Increasing these corrections within the explored parameter range leads to improved performance in reproducing the present-day ice-sheet state and the temporal evolution of

the ice margin. In the case of b, this improvement levels off in the upper part of the range, and then decrease slightly.

The precipitation scaling factor f_p, the other parameter directly involved in SMB, shows no statistically significant relationship with any of the evaluated scores. This is not surprising, as f_p is only varied within the northeastern basin and therefore has a limited impact on the broader ice-sheet geometry. This result should therefore not be interpreted as an indication that precipitation is unimportant for the deglacial evolution of the GrIS: it only reflects the limited spatial extent over which f_p is perturbed in this ensemble. A precipitation scaling applied over the whole ice sheet would likely have a larger impact on the scores.

L412-L415

Regarding other climate-forcing parameters, the ocean sensitivity κ shows a comparatively large impact on simulation performance. Higher values of κ improve the ice-cover score for both the present day ($S_{ice\ cover}$) and, to a lesser extent, its temporal evolution throughout the deglaciation (S_{iso}). Furthermore, κ alone explains 59% of the variance of the present-day ice surface velocity score (S_u). S_u increases with κ and levels off for $\kappa \gtrsim 8$ m yr^{-1} K^{-1}

L422-426

The parameters associated with basal dynamics (q, z_0, and δ) as well as the ice rheology parameter (E_{shear}) affect the evaluated scores differently. The ice thickness score (S_H), and consequently the bedrock-elevation score (S_{zbed}) – which is largely controlled by the ice load above – together with S_{DYE-3}, decrease with increasing δ, whereas they tend to increase with higher values of z_0, q, and E_{shear}. The influence of E_{shear} is however generally weak across the evaluated scores.

L434-L435

The opposite behavior is observed for the present-day ice surface velocity score (S_u) and the surface elevation evolution at GRIP (S_{GRIP}), both of which favor parameter combinations yielding slower ice dynamics and reduced deformation. In the case of S_{GRIP}, the relationship with δ (R^2 = 0.18) is non-linear : the score drops sharply towards the lower end of the explored range and remains nearly constant for $\delta \gtrsim 0.03$.

L442-L443

The calving parameter κ_t has a limited impact on most scores, with the notable exception of the LGM ice mask score (S_{LGM}), for which it explains 56% of the variance through a negative linear relationship.

L451-L455

[...] and the asthenosphere is viscous with a relaxation time τ. Both parameters explain only a small fraction of the variance of any score ($R^2 \leq 0.06$). The largest values are found for H_e with S_{LGM}, through a negative relationship suggesting that an accurate representation of the LGM extent requires a relatively thin lithosphere, and for τ with S_GRIP, favoring short relaxation times, [...]

L457-L459

Finally, the parameters controlling basal friction in the NEGIS have a very limited impact: f_{low} shows no statistically significant relationship with any score, and f_{mid} explains only 2% of the variance of S. Surprisingly, no parameter alone explains more than 3% of the variance of S_{NGRIP}, which shows only weak non-linear dependences on δ and f_{mid}.

3) To avoid increasing the length of the paper further, we will add the description of the method to the Supplement, together with the individual scatter plots and fitted curves for the 13 parameter-score pairs with $R^2 > 0.15$ (Fig. 2 below). The supplement section would be as follows:

Polynomial regression analysis of score-parameter relationships

To characterise the relationship between each score and each perturbed parameter, including possible non-linear dependencies, for each parameter-score pair, we fitted linear (d=1), quadratic (d=2) and cubic (d=3) polynomials and selected the most appropriate one using the Bayesian Information Criterion (BIC, Neath and Cavanaugh, 2012; Schwarz, 1978):

$$BIC = n \ln\left(\frac{RSS}{n}\right) + k \ln(n),$$

where RSS is the residual sum of squares, k=d+1, and n is the number of ensemble members. The BIC penalises model complexity, so a higher polynomial degree is only selected when it improves the fit sufficiently.

The strength of each relationship is quantified by the coefficient of determination of the selected fit, $R^2 = 1 - RSS/TSS$, where TSS is the total sum of squares of the score. R^2 is the proportion of the variance of the score explained by the parameter alone. The statistical significance of each fit was assessed with an F-test (Draper and Smith, 1998), and a Bonferroni correction (Dunn, 1961) was applied for the 130 tests, so that fits are considered significant if $p < 0.05/130 \approx 3.8 \times 10^{-4}$. Because of the large ensemble size, even very weak relationships are statistically significant. R^2 is therefore the most informative quantity.

The BIC selects a linear model for 100 of the 130 pairs, a quadratic model for 21 and a cubic model for 9. For the non-linear pairs, the gain in explained variance with respect to the linear fit exceeds 0.05 in only four cases: S_{iso}-b, S_{ice} cover-b, S_H-b and S_{GRIP}-δ (gains of 0.07–0.18). Figure S2 shows the individual scatter plots and fitted curves for the 13 pairs with $R^2 > 0.15$.

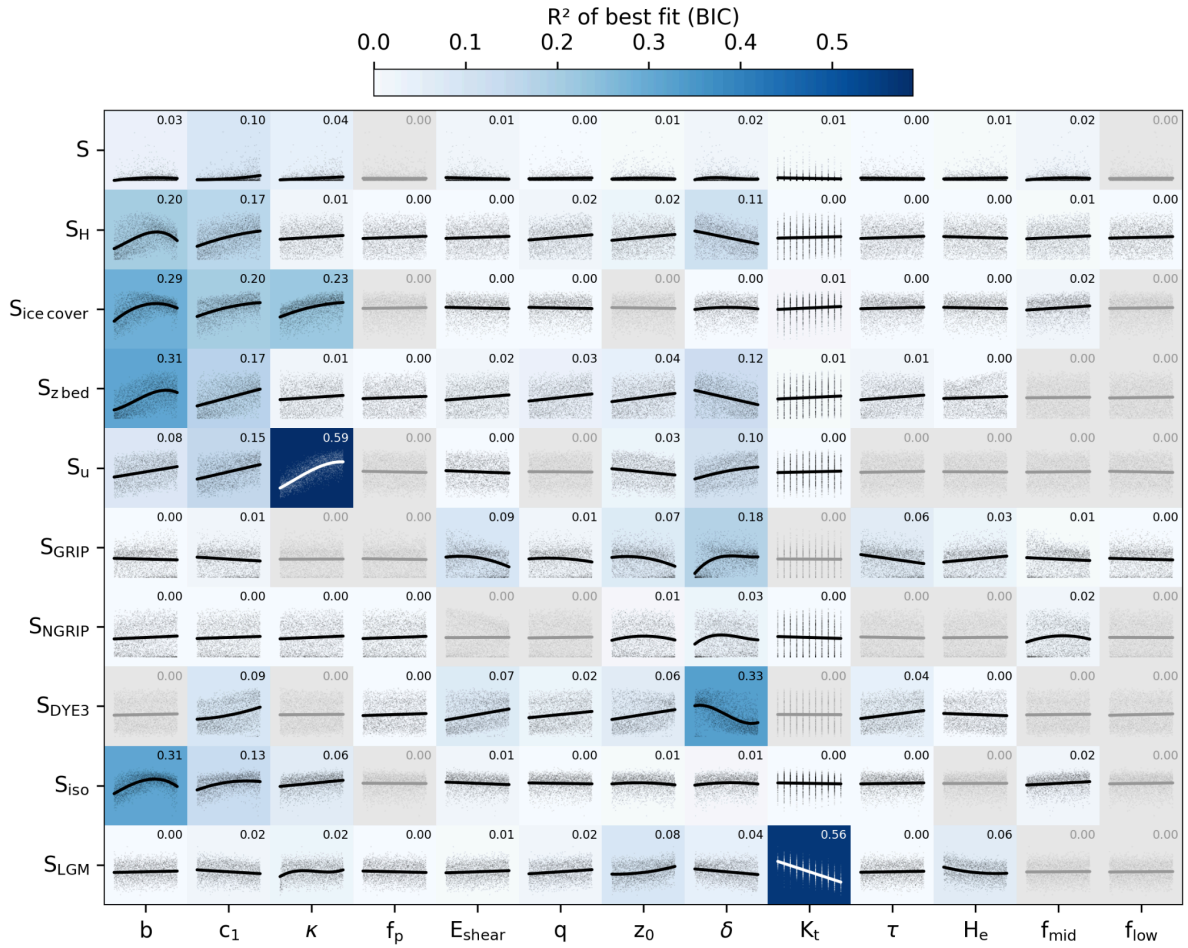


Figure 1: Proportion of explained variance R^2 of the best polynomial fit between each pair score-parameter across the 3000-member ensemble. For each pair, the degree of the polynomial (linear, quadratic or cubic) was selected with the Bayesian Information Criterion (BIC). Each cell shows the selected fit (curve) and the individual simulations (small dots). Cell colours indicate R^2, and the number in each cell gives its value. The horizontal axis of each curve spans the sampled range of the parameter, and the vertical axis spans the range of the corresponding score across the ensemble (common to all cells in a row). Grey cells indicate fits that are not statistically significant at the 5 % level after a Bonferroni correction for the 130 tests ($p \geq 3.8 \times 10^{-4}$).

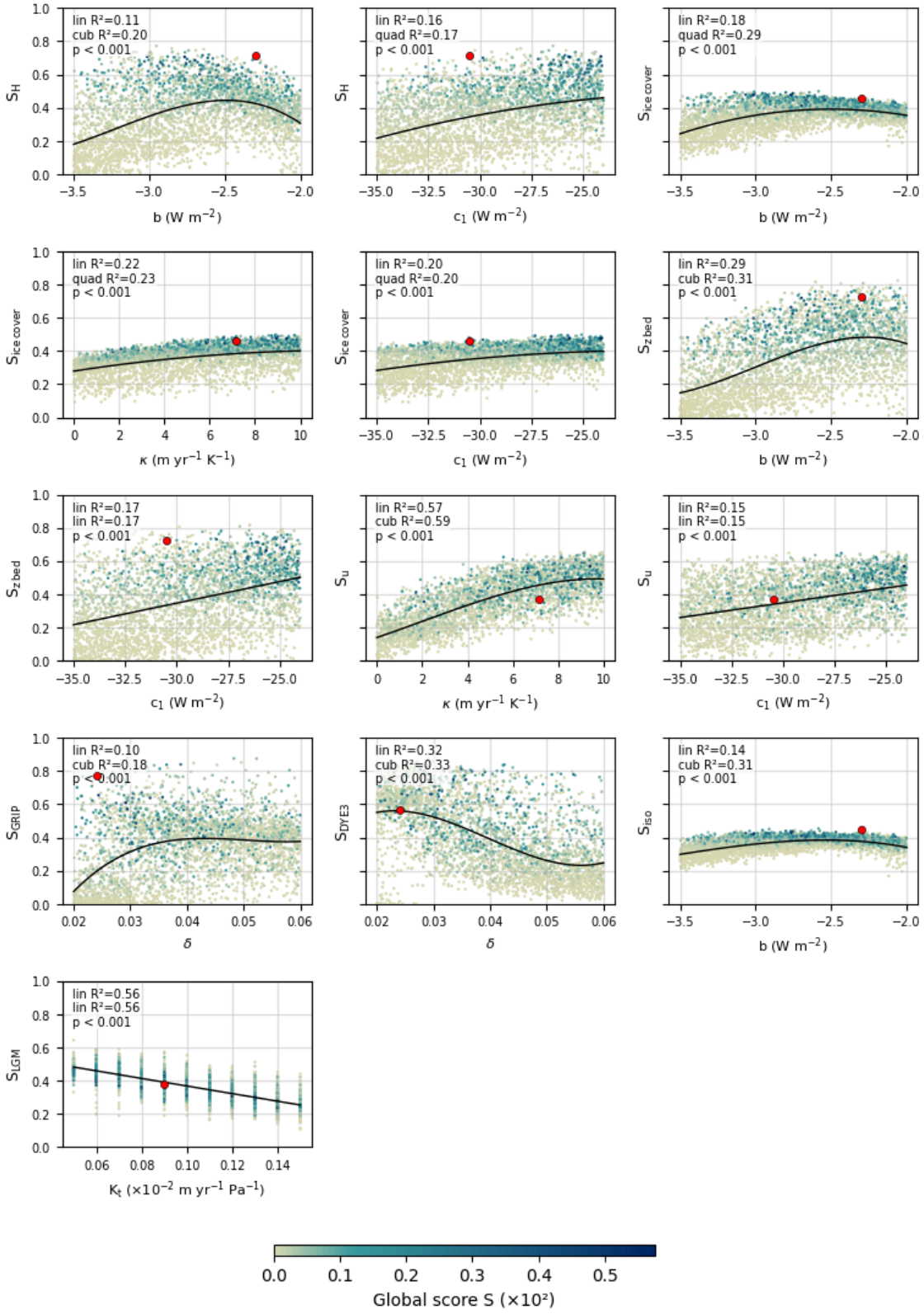


Figure 2: Scatter plot between model parameters and scores for the 13 parameter-score pairs with R^2 > 0.15 for the best-fitting model. Each dot is one ensemble member, coloured by its global score. The red dot marks the best simulation. The black line shows the best polynomial fit selected with the BIC. P-values are for the best fit (F-test) after Bonferroni correction for the 130 score-parameter pairs.

[SHAP Analysis]

We thank the reviewer for this comment. However, we respectfully disagree that this section does not add to the paper. The emulator addresses a question that the previous analysis cannot: the combined effect of the parameters on the overall performance of the simulations. While the analysis of Fig. 1 only considers the relationship between each parameter and each score individually, SHAP values quantify the total contribution of each parameter to the global score, including both its direct effect and its interactions with the other parameters. We would therefore like to keep this section, while moving the description of the methodology to the Methods section, making its contribution more explicit and making the analysis more robust, as explained below.

Prompted by this comment, we have nevertheless improved the emulator. Since S is defined as a product of exponentials, it spans more than 20 orders of magnitude across the ensemble. As a result, the original emulator was dominated by the simulations with the highest scores, while differences among poorly performing simulations, all with $S \approx 0$, were essentially invisible to it. The emulator is now trained on ln(S), which is equal to the negative total normalised misfit ($\ln S^j = -\sum_i M'^j_i$) and therefore has a direct physical interpretation. The revised emulator explains 82% of the variance (cross-validated $R^2 = 0.82$; $R^2 = 0.83$ on an independent test set). Importantly, the individual parameters together explain only about 24% of the variance of ln S when their effects are considered separately (sum of the R^2 of the best polynomial fits between ln(S) and each parameter, computed as in Fig. 1). This shows that the performance of the simulations is largely controlled by combinations of parameters, which cannot be captured by a pairwise analysis.

With the revised emulator, the three parameters that were previously identified as the most influential, namely the climate-forcing parameters c_1, b and kappa, remain among the four most important ones, with the same direction of influence (Fig. 3). However, they are now preceded by δ, which shows a long negative tail for low values. This indicates that low values of δ, largely in combination with other parameters rather than on their own, strongly degrade the performance of the simulations, reducing S by several orders of magnitude in some cases. This effect was not apparent in the pairwise analysis (Fig. 1, and Fig. 11 of the original manuscript): although δ explains a substantial fraction of the variance of several individual scores (e.g. $R^2 = 0.33$ for S_DYE3 and 0.18 for S_GRIP), its R^2 with the global score S is small (0.02).

Additional differences with respect to the original analysis, which will be included in the revised text, concern z_0 and τ, both of which can degrade the performance for high values ($z_0 > -250$ m and $\tau > 2000$ yr). Since lower values of δ and higher values of z_0 reduce basal friction, these results show that an excessively low basal friction, in combination with other parameters, degrades the performance of the simulations. For τ, the results indicate that slow isostatic adjustment is detrimental, consistent with the preference for short relaxation times found for S_GRIP in the pairwise analysis. The threshold behaviour of f_mid (f_mid < 0.45) is confirmed, whereas the statement on the precipitation scaling factor f_p has been removed, as it is not supported by the revised emulator.

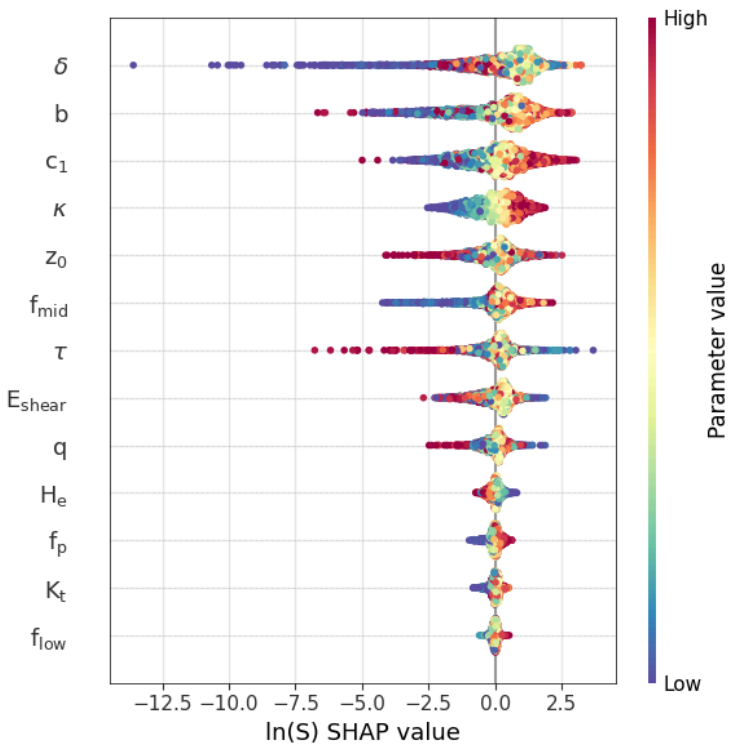


Figure 3: SHAP values of ln(S) for each of the parameters perturbed in the ensemble, calculated with a XGBRegressor model trained on ln(S) (R^2 = 0.83 on the independent test set). Positive (negative) values indicate that the parameter value increases (decreases) the predicted ln(S).

In summary, the changes to the manuscript are:

1) The description of the emulator and the SHAP methodology (L461–478) has been moved to a new Sect. 2.4, "Emulator-based sensitivity analysis", placed after L240 (end of Sect. 2.3). The emulator is now trained on ln(S) instead of S, and the performance metrics have been updated accordingly (cross-validated R^2 = 0.82, R^2 = 0.83 on the independent test set):

2.4 Emulator-based sensitivity analysis

To explore the sensitivity of the overall performance of the simulations to the ensemble parameters, we developed an emulator model based on the eXtreme Gradient Boosting (XGBoost) algorithm (Chen et al., 2016). The model uses the specific parameter combinations of each simulation as input features to predict the logarithm of the global score, ln(S). Since S is defined as a product of exponentials (Eq. 16), it spans more than 20 orders of magnitude across the ensemble, so that an emulator trained on S is dominated by the few best-performing simulations, while differences among poorly performing simulations (all with $S \approx 0$) are essentially invisible to it. In contrast, $\ln(S) = -\Sigma_i M'_i$ is the negative total normalised misfit, which varies smoothly across the ensemble and has a direct physical interpretation. The dataset was partitioned into a training set (90%) and an independent testing set (10%). Hyperparameter optimization was conducted via a 5-fold cross-validated grid search on the training partition. To mitigate overfitting within this high-dimensional parameter space, we enforced regularization constraints, including a penalty (λ term in Chen et al., 2016) and a restricted maximum tree depth. The optimized emulator achieved a cross-validated determination

coefficient (R^2) of 0.82. Following the selection of optimal hyperparameters, the model was retrained on the full training set, yielding an R^2 of 0.99. The final model performance was evaluated on the independent testing set, resulting in an R^2 of 0.83. The close agreement between the cross-validation and testing metrics confirms that the model successfully generalizes the primary physical sensitivities without overfitting to the training data.

Thereafter, we performed an interpretability analysis to attribute parameter influence to ln(S) within the fraction of variance captured by the XGBoost model. For this purpose, we employed the SHAP (SHapley Additive exPlanations) framework, based on the Shapley values developed in the game-theory approach that quantifies the marginal contribution of each parameter to the predicted score (Shapley et al., 1953). This approach allows for the quantification of the relative importance of parameters in the global performance of the simulations (Lundberg et al., 2020, Kellner et al., 2022), accounting for both the direct effect of each parameter and its interactions with the other parameters.

2) Fig. 12 has been replaced by the SHAP summary plot of the revised emulator, and the analysis in Sect. 3.3.1 (L479–501) has been updated accordingly:

3.3.1 Emulator-based sensitivity analysis

The SHAP values obtained with the emulator described in Sect. 2.4 are summarised in Fig. 12 (all the SHAP values are shown against the different parameter values in Fig. S1 of the Supplementary Material), illustrating both the hierarchy of parameter importance and the directionality of their impact. Within this plot, the vertical position indicates the relative importance of the variable, while the horizontal axis represents the magnitude and sign of its influence on ln (S). Positive SHAP values indicate an improvement in performance (lower total misfit), and a SHAP value of -2.3 corresponds to a reduction of S by one order of magnitude. For instance, c1 is one of the most influential parameters: low values of c1 are associated with lower scores, whereas high values (shown in red) yield SHAP values of up to about 3, indicating moderate-to-high scores. In other words, low values of c1 consistently degrade S, while high values mostly improve it, with a marked improvement in score for $c_1 > -27.5 Wm^{-2}$. As indicated by Eq. 6, this suggests that the prescribed precipitation (Badgeley et al., 2020} and temperature (Buizert et al., 2018) forcings for this setting need an upward adjustment to surface melting to achieve satisfactory model performance.

The SHAP value analysis is complementary to the analysis presented in the previous section in several ways: (1) it provides an independent method that strengthens conclusions shared by both approaches, and (2) it accounts for interactions between parameters, which cannot be captured by the pairwise analysis of Fig. 11. This is particularly relevant here, since the individual parameters together explain only about 24% of the variance of ln(S) when considered separately (sum of the R^2 of the best polynomial fits, computed as in Fig. 11), whereas the emulator explains 82%.

As revealed by Fig. 11, the climate-forcing parameters c_1, b and κ are among the most influential, with the same sense of influence as in the pairwise analysis: higher values of c_1 and κ improve performance, whereas b improves it up to $b \approx -2.3 Wm^{-2}$, with a degradation towards the upper end of the explored range. However, the parameter with the largest impact is δ, which shows a long negative tail for low values ($\delta \lesssim 0.025$): in combination with other parameters, these values reduce S

by up to several orders of magnitude (SHAP values down to about -13). This effect is hidden in Fig. 11 because S compresses all poorly performing simulations to S ≈ 0. Another parameter whose impact is not apparent in Fig. 11 is f_{mid}. This variable shows a threshold behavior: when f_{mid} takes low values within the explored range (f_{mid} < 0.45), basal friction in the NEGIS central stream is reduced excessively, allowing an unrealistically large advance of the ice margin in that region at the present day.

Additionally, z_0 and the asthenospheric relaxation time τ also show a significant non-linear impact: their influence is neutral over most of the explored range, but becomes detrimental for $z_0 \gtrsim$-250m and $\tau \gtrsim$ 2000 yr. Since lower values of δ and higher values of z_0 both reduce basal friction (Eqs. 2–4), these results indicate that excessively low basal friction degrades the performance of the simulations. For τ, they indicate that slow isostatic adjustment is detrimental, consistent with the preference for short relaxation times found for S_{GRIP} (Fig. 11). E_{shear} shows an intermediate optimum around 1.2–1.3, with both lower and higher values slightly degrading performance. Overall, the SHAP results are consistent with the preceding analysis, while providing a more thorough characterisation of parameter influence, particularly for parameters exhibiting threshold-dependent behaviour and interactions with other parameters.

3) Replace Figure S1 with the updated version (Figure 4 in this text).

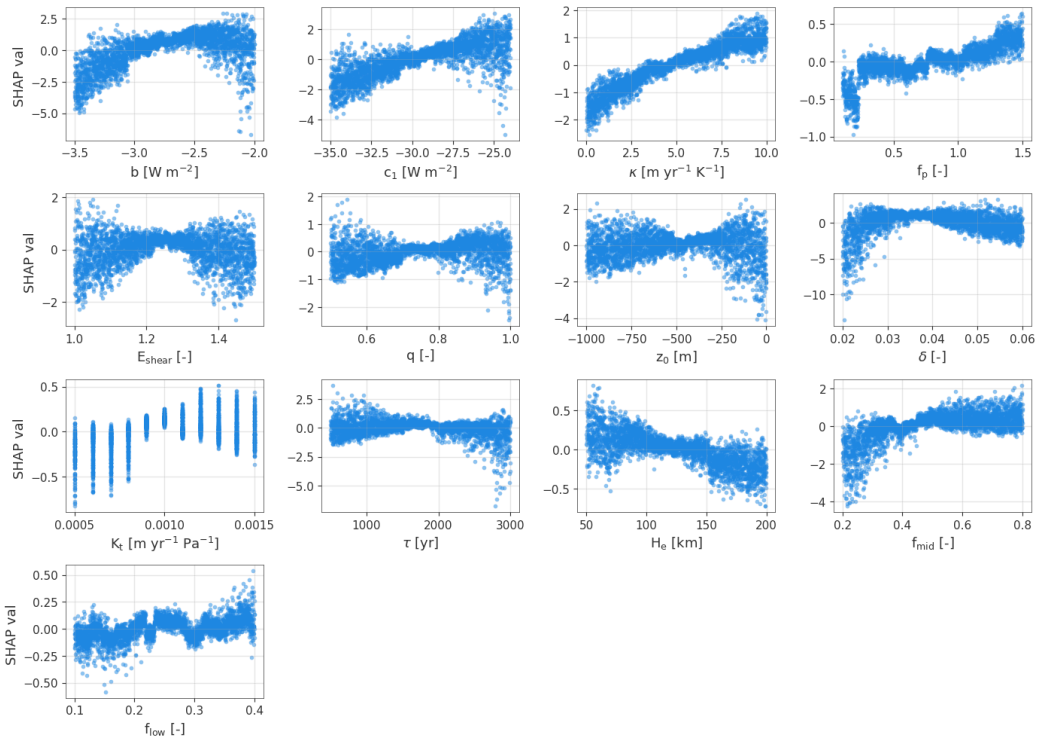


Figure 4: SHAP scatter plots showing the relationship between individual parameter values and their contribution to ln(S). Positive SHAP values indicate an increase in the score, while negative values indicate a decrease.

4) We have modified the Discussion and Conclusions accordingly to account for the influence of δ:

L563-L564

Finally, our results show that, for the forcing and parameter space considered in this study, the performance of the simulations in reproducing deglacial observables is primarily driven by the specific ice-sheet sensitivity to climate forcing, together with basal friction, whose lowest values strongly degrade the overall performance (Sect. 3.3.1).

L650-L651

The sensitivity of the ice sheet to atmospheric and oceanic climate forcing is the primary control on the last deglaciation performance, while excessively low basal friction can strongly degrade it.

Fp – I find it confusing that this parameter is only tuned in one basin. It's hard to keep track of this when the info is not in Table 2 and not called out in later figures. I have two concerns: 1) the paper gives a false sense that precipitation scaling isn't important and 2) there are other important details of other parameters that I didn't catch and should be highlighted in Table 2.

We agree with the reviewer. In the revised manuscript, we will indicate in Table 2 and in the text that f_p is only perturbed in the northeastern basin. We will also add a column to Table 2 specifying the spatial domain of each parameter. This will also make explicit that f_mid and f_low only apply to the NEGIS region, while the remaining parameters apply to the whole ice sheet. We will revise the text to avoid giving the impression that precipitation scaling is unimportant: its limited effect on the scores reflects the restricted area where it is perturbed.

Therefore, the changes will be:

1) In Sect 3.3, already taken into account in the previous answer in L403-L411.
2) In Table 2, where the column "Domain" will be added:

Parameter	Description	Domain	Units	Range
Ice dynamics				
f_{mid}	Mid NEGIS basal friction coefficient	NEGIS mask	–	[0.2, 0.8]
f_{low}	Low NEGIS basal friction coefficient	NEGIS mask	–	[0.1, 0.4]
q	Friction law exponent	Ice sheet	–	[0.5, 1]
z_0	Bedrock e-folding depth for basal friction reduction	Ice sheet	m	[-1000, 0]
δ	Fraction of overburden pressure	Ice sheet	–	[0.02, 0.06]
E_{shear}	Enhancement factor for the shearing-driven areas	Ice sheet	–	[1.0, 1.5]
κ_{T}	Calving scaling parameter	Ice sheet	$\mathrm{m\,yr^{-1}\,Pa^{-1}}$	$[5.0{\cdot}10^{-4}, 1.5{\cdot}10^{-3}]$
Atmosphere				
b	Latitude correction for surface melt	Ice sheet	$\mathrm{W\,m^{-2}}$	[-3.5, -2.0]
c_1	Additional correction for surface melt	Ice sheet	$\mathrm{W\,m^{-2}}$	[-35, -24]
f_{p}	Precipitation scaling factor	Northeastern basin	-	[0.1, 1.5]
Ocean				
κ	Ocean sensitivity	Ice sheet	$\mathrm{m\,yr^{-1}\,K^{-1}}$	[0, 10]
Solid Earth				
H_{e}	Elastic lithosphere thickness	Ice sheet	km	[50, 200]
τ	Asthenosphere relaxation time	Ice sheet	yr	[500, 3000]

Weighting of Model Scores – I think there needs to be more discussion of how to combine the different model scores. The present day ones make sense to me overall. It's the paleo ones that I'm a bit concerned about. Of the five paleo scores, there are ice sheet thinning history. Of these three, two are essentially the same (GRIP and NGRIP) which is noted in the paper, but we know a priori they are going to be incredibly similar given their proximity at the summit of the ice sheet. The other, Dye3, is wildly more uncertain. It's easy to forget that Dye3 is not at a divide, so the ice has flowed through strong climate gradients, and the LGM ice is deep and highly thinned. I feel like having these three be separate adds too much weight to them. It seems worth testing how different weighting of the scores going into the total score affects the best model. Looking at Figure 9, it seems possible that you could get a big shift in the parameters for the best fitting model.

We thank the reviewer for raising this point, which we have now tested explicitly. We recomputed the global score S under an alternative weighting in which the three ice-core elevation scores (S_NGRIP, S_GRIP and S_DYE-3) share the weight of a single observable (1/3 each), and compared the results with those obtained using the original equal weighting (Table 1). Under the 1/3 weighting, our best simulation (#2572) ranks 2nd out of 3,000 members, with member #1081 in first place. The two members are nearly equivalent: their global scores differ by only ~10%, and #1081 ranks 3rd under the original weighting.

Weighting	**Weights (GRIP / NGRIP / DYE-3)**	**Best simulation**	**Rank of #2572**	**Top 3**
Original	1 / 1 / 1	2572	1	2572, 2276, 1081

Ice cores as one observable	1/3 / 1/3 / 1/3	1081	2	1081, 2572, 2276

Table 1: Summary of the weighting results.

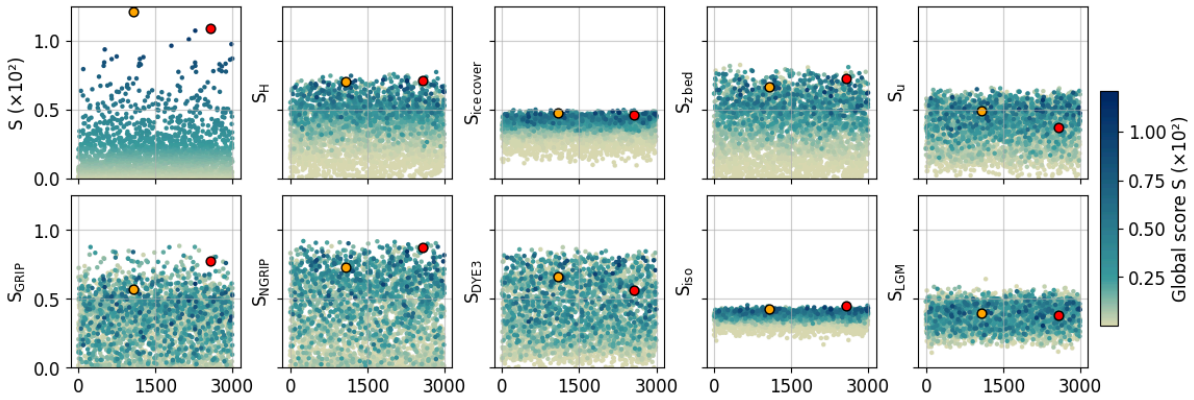


Figure 5: Global score S under the 1/3 weighting and individual scores for each of the 3,000 simulations. Red and orange dots indicate simulations #2572 and #1081, respectively. The colour bar shows the global score S, and the x-axis shows the simulation number. This figure is analogous to Fig. 9 of the manuscript; the only differences are the first panel (S) and the colour bar, both computed with the 1/3 weighting, and the addition of simulation #1081 in orange.

The two simulations perform very similarly across all metrics. The main differences are:

- For the present day, #1081 performs slightly worse in ice thickness (RMSE ~10 m higher, less than 10% increase) but slightly better in surface velocity (RMSE ~10 m yr^{-1} lower, also less than 10% decrease).
- #1081 yields sea-level contributions of −5.44 m at the LGM and +0.62 m after the HTM, compared with −5.75 m and +0.45 m for #2572. Both lie within our 68% ranges, which are identical under the original and 1/3 weighting ([−6.31, −3.93] m at the LGM and [0.43, 1.18] m after the HTM, calculated by ranking simulations according to their scores), while the 95% ranges change by at most 0.06 m.
- #1081 simulates a larger post-HTM retreat in the Kangerlussuaq region (maximum value ~136 km vs ~120 km), a region where our best simulation already overestimates the retreat relative to the 2–26 km suggested by geological reconstructions (Sect. 4.2 of the manuscript).
- #1081 performs worse at GRIP and NGRIP but better at DYE-3 (Fig. 5).
- There are no significant differences in the LGM state or in the fit to the PaleoGrIS isochrones.

Although the two simulations perform similarly, some of their parameters differ substantially (e.g. c_1, q and z_0, by 40-56 % of their sampled range). Good performance thus arises from combinations of parameters that compensate each other, not from a unique configuration. Our analysis therefore identifies parameter ranges that tend to improve or degrade performance, rather than a single optimal set.

Since #1081 does not improve the overall fit to present-day observations (it improves surface velocities but worsens ice thickness), or to the past constraints, we consider it more appropriate to retain our original best simulation.

Regarding the individual ice cores, we agree that S_GRIP and S_NGRIP are correlated (Fig. 10 of the manuscript). However, they are not redundant: (i) they show different parameter sensitivities (Fig. 11 of the manuscript), with S_GRIP being sensitive to E_shear and τ, whereas S_NGRIP shows no clear linear dependence on any individual parameter and is mainly linked to the dynamics of the NEGIS (Tabone et al., 2024); and (ii) their reconstructed elevation histories differ through time (Fig. 4 of the manuscript).

It is also true that the DYE-3 record is more uncertain, since the site is not located at an ice divide and the ice has flowed through strong climatic gradients, as the reviewer mentions. We will state this explicitly in the revised manuscript. Nevertheless, DYE-3 remains the only elevation constraint in southern Greenland, and its score correlates with the present-day and LGM scores (Fig. 10 of the manuscript), indicating that it is consistent with the other constraints. Moreover, our best simulation is not driven by this site: its S_DYE-3 (0.56) is the lowest of its three ice-core scores.

For these reasons, we propose to maintain equal weights for all observables, as the least subjective choice, together with our original best simulation. In the revised manuscript, we will justify this weighting choice in the Methods section.

In summary, the modifications to the manuscript will be:

1) L182 (at the end of the sentence) we will add:

Note that the DYE-3 record is subject to larger uncertainties than those of the central ice-core sites, since DYE-3 is not located at an ice divide, and the ice at this location has flowed through strong climatic gradients (Vinther et al., 2009).

2) L238 we will add:

[...]the total normalized misfit. In this way, all observables are equally weighted in the global score S. We consider this the least subjective choice, as any alternative weighting would require prioritising some constraints over others. We nevertheless assessed the robustness of this choice with respect to the weight given to the ice-core records, motivated by the larger uncertainty of the DYE-3 record and the proximity of the GRIP and NGRIP sites. To this end, we recomputed S with the three ice-core scores sharing the weight of a single observable (1/3 each), and found that this alternative weighting does not substantially change the results (Supplementary Material, Sect. X).

Simulations with S^j closer to 1 provide a better overall [...]

3) We will add a section to the supplementary material with Table 1, Fig. 5 and the following text:

In the main text, all observables are equally weighted in the global score S (Sect. 2.3). Here, we assess the robustness of this choice with respect to the weight given to the ice-core records. Two

arguments could justify reducing this weight: the GRIP and NGRIP sites are located close to each other in the central part of the ice sheet, and their scores are correlated (Fig. 10); and the DYE-3 record is subject to larger uncertainties, since the site is not located at an ice divide (Vinther et al., 2009). We therefore recomputed S with the three ice-core scores (S_GRIP, S_NGRIP and S_DYE-3) sharing the weight of a single observable (1/3 each).

The results are summarised in Table 1. Under this alternative weighting, the best-fit simulation of the main text (#2572) ranks second out of 3,000 members, with member #1081 in first place (Fig. S3). The global scores of both members differ by only ~10 %, and #1081 ranks third under the original weighting. The two simulations perform very similarly across all metrics. The main differences between them are:

- Present day. Simulation #1081 has an ice thickness RMSE ~10 m higher (an increase of less than 10 %) and a surface velocity RMSE ~10 m yr^{-1} lower (a decrease of less than 10 %).
- Ice cores. Simulation #1081 performs worse at GRIP and NGRIP but better at DYE-3. (Fig. S3)
- Sea-level contributions. Simulation #1081 yields contributions of −5.44 m at the LGM and +0.62 m after the HTM, compared with −5.75 m and +0.45 m for #2572. Both lie within the 68 % ranges of the ensemble, which are identical under both weightings ([−6.31, −3.93] m at the LGM and [0.43, 1.18] m after the HTM), while the 95 % ranges change by at most 0.06 m.
- Kangerlussuaq. Simulation #1081 simulates a larger post-HTM retreat in the Kangerlussuaq region (maximum of ~136 km vs ~120 km), a region where the best-fit simulation already overestimates the retreat relative to geological reconstructions (Sect. 4.2).
- No significant differences are found in the LGM state or in the fit to the PaleoGrIS isochrones.

Figures 2-4: I'm unclear if "best simulation" is referring to the overall best, or the best for that sub-score. Can you clarify that in the captions and text.

"Best simulation" refers to the overall best simulation, i.e. the one with the highest global score S. We will clarify this in the captions of Figs. 2-4 and at the beginning of the Results section.

In the captions of Figures 2-4 we will apply these changes:

1) In Figure 2: the overall best-scoring simulation (black line)
2) In Figure 3: the overall best simulation is marked in black
3) In Figure 4: the overall best simulation in black
4) In L251:

Although all metrics are weighted equally in the global score S (Sect. 2.3), the best simulation (black line; i.e. the one with the highest global score S), shows excellent agreement with the observed grounded area from Leger et al. (2024), [...]

References

Badgeley, J. A., Steig, E. J., Hakim, G. J., and Fudge, T. J.: Greenland temperature and precipitation over the last 20,000 years using data assimilation, Climate of the Past, 16, 1325–1346, https://doi.org/10.5194/cp-16-1325-2020, 2020.

Buizert, C., Keisling, B. A., Box, J. E., He, F., Carlson, A. E., Sinclair, G., and DeConto, R. M.: Greenland-wide seasonal temperatures during the last deglaciation, Geophysical Research Letters, 45, 1905–1914, https://doi.org/10.1002/2017GL075601, 2018.

Cuzzone, J., Briner, J., Otiniano, G., et al.: Greenland Ice Sheet to fall below its Holocene minimum size by 2300, Research Square [preprint], https://doi.org/10.21203/rs.3.rs-9962320/v1, 29 June 2026.

Goelzer, H., Nowicki, S., Payne, A., Larour, E., Seroussi, H., Lipscomb, W. H., Gregory, J., Abe-Ouchi, A., Shepherd, A., Simon, E., Agosta, C., Alexander, P., Aschwanden, A., Barthel, A., Calov, R., Chambers, C., Choi, Y., Cuzzone, J., Dumas, C., Edwards, T., Felikson, D., Fettweis, X., Golledge, N. R., Greve, R., Humbert, A., Huybrechts, P., Le clec'h, S., Lee, V., Leguy, G., Little, C., Lowry, D. P., Morlighem, M., Nias, I., Quiquet, A., Rückamp, M., Schlegel, N.-J., Slater, D. A., Smith, R. S., Straneo, F., Tarasov, L., van de Wal, R., and van den Broeke, M.: The future sea-level contribution of the Greenland ice sheet: a multi-model ensemble study of ISMIP6, The Cryosphere, 14, 3071–3096, https://doi.org/10.5194/tc-14-3071-2020 , 2020.

Neath, A. A. and Cavanaugh, J. E.: The Bayesian information criterion: background, derivation, and applications, WIREs Computational Statistics, 4, 199–203, https://doi.org/10.1002/wics.199, 2012.

Schwarz, G.: Estimating the dimension of a model, The Annals of Statistics, 6, 461–464, https://doi.org/10.1214/aos/1176344136, 1978.

Tabone, I., Robinson, A., Montoya, M. *et al.* Holocene thinning in central Greenland controlled by the Northeast Greenland Ice Stream. *Nat Commun* **15**, 6434 (2024). https://doi.org/10.1038/s41467-024-50772-5

Draper, N. R. and Smith, H.: Applied Regression Analysis, 3rd edn., Wiley, New York, https://doi.org/10.1002/9781118625590 , 1998.

Dunn, O. J.: Multiple comparisons among means, J. Am. Stat. Assoc., 56, 52–64, https://doi.org/10.1080/01621459.1961.10482090 , 1961.

Vinther, B. M., et al.: Holocene thinning of the Greenland ice sheet, Nature, 461, 385–388, 2009.