

Response Letter

We would like to thank the editor and two reviewers for their insightful comments and suggestions concerning our manuscript, *Future learning and uncertainty reductions in projections of the Amery Ice Shelf catchment, Antarctica*. In the following, we respond to each of the reviewers' comments in detail. The reviewer summary and comments are listed in *italic*, while our responses are highlighted in blue and not italicized. To aid further review, we will mark all new additions/revisions to our previous manuscript in blue, except when we intend to add entire section/subsections. In those instances, we will indicate the subsection heading in blue within the revised manuscript.

We wish to highlight at the outset that, in the process of responding to Reviewer 1's General Comment 2 (also raised by Reviewer 2), we determined that we needed to make two methodological changes accounting for (i) temporal correlation and (ii) exclusion of emulator code uncertainty. These changes affect all the results and subsequent discussion, and those sections will be substantially revised in the revised manuscript. We describe these changes in detail in our response to Reviewer 1's General Comment 2, and we highlight in other responses where results are likely to change. We have already recalculated all results, and while many of the details change, the big picture conclusions are largely unchanged.

We also note that the author order will be updated in the revised manuscript, as previously communicated with the editorial office and agreed upon by all authors.

Response to Reviewer 1

Review summary. *This manuscript covers an approach to determine how future observations of ice-sheet loss could change parametric and forecast uncertainty as new observations are made. The analysis focuses on Lambert glacier/Amery ice shelf. An emulator based on a MALI model ensemble is generated and used for Bayesian calibration of future observations to revise assessments of a small number of scalar parameters.*

I really like the idea of this manuscript. The effect of future observations is not something generally considered in UQ. The authors are comprehensive by considering a range of future observation trajectories; and the discussion is insightful regarding the results and their physical interpretations.

General comments

General comment 1 (complexity and approachability). *In terms of general comments, one aspect I will comment on is the complexity of the approach and making it approachable. A GP emulator must be defined; a set of synthetic trajectories must be generated; and for each, Bayesian posteriors must be performed at 30 different time instances, for 3 different horizons. When discussing results, I think that methods could do with more systematic descriptions, and improved methods of referring to quantities could be found. In my detailed comments I give some examples. Also, I'm not sure if there is a standard style in the journal for different mathematical objects (scalars, vectors, functions, tensors, etc.) — if not, I suggest the authors create their own. For instance, I was unsure whether the named parameters were scalars or vectors until further down.*

We will restructure Section 2 to make the multi-step workflow more approachable. Specifically, we will open Section 2 with a concise overview of the full procedure before describing each step in detail:

- (i) **Emulator construction:** train scalar GP emulators that map the six (scalar) MALI parameters to the cumulative grounded mass change at each 15-year analysis year, and adopt the PCA-GP emulator of Jantre et al. (2024) for the volume-above-flotation (VAF) change (transformed into sea-level contribution) time series;

- (ii) **Synthetic future-observation trajectory generation:** draw 100 parameter vectors from the present-day (year 2015) posterior and, for each, generate one synthetic future-observation trajectory of cumulative grounded mass change with temporally correlated observational noise (see our response to your General comment 2);
- (iii) **Sequential Bayesian calibration:** for each trajectory and analysis year t , recalibrate the parameters using all observations available through t (MCMC using the NUTS sampler);
- (iv) **Forward uncertainty propagation:** propagate the recalibrated parameter posterior samples through the PCA–GP emulator of VAF change to obtain sea-level projections at 2100, 2200, and 2300;
- (v) **Quantifying future learning:** measure learning as contraction of 90% credible-interval widths and summarize the results across the 100 trajectories.

To clarify the design dimensions (the review refers to “30 different time instances”), we perform sequential calibrations at 19 analysis years (every 15 years, 2030–2300) for each of 100 synthetic trajectories, giving 1900 calibrations in total, and evaluate sea-level-contribution learning at three projection horizons (2100, 2200, 2300). We will also state an explicit notation convention at the start of Section 2: scalars in lightface italic (e.g., q , γ_0) and vectors in lowercase boldface (e.g., θ), and clarify at first mention that all six model parameters are scalars. The specific passages flagged by the reviewer are addressed below; reorganizing Section 2 around the workflow above will also remove the apparent repetition the reviewer noted.

General comment 2 (Major revision – temporal correlation of cumulative observations). *There is another aspect that I think important enough to make general. The calibrations of future observations use an error model that treats cumulative mass loss at various horizons as independent for reasons of tractability (I assume). But allowing for correlation, I do not believe is intractable, and furthermore I believe disallowing it introduces zero-order errors in the results, to the point of near logical contradiction — I make specific comments about this below. The ignoring of correlations is brought up as a limitation — but simply saying the omission of something is a limitation, when it is as important, and easy NOT to omit, is not enough, and I am hardly the only person who would notice this and think this. The authors need to account for these correlations in their analysis or give a rigorous explanation in-manuscript of why it cannot be done, and what errors it introduces. I am selecting “Major revisions” for this reason, though if I am mistaken or misunderstanding in some way of which I can be convinced, then major revisions may not be necessary.*

The reviewer is correct, and we thank them for pressing on this point: the observational errors in the cumulative grounded mass-change observations are temporally correlated *by construction*. Treating them as independent in the likelihood, as was done in our submitted manuscript, mis-specifies the observation error model. We now account for this using a temporally correlated observation error model, and our revised manuscript will use this corrected error model. Below we describe analysis we have done to evaluate the impact this change will have on the manuscript. In that analysis, we identified a second design choice that was masking the effect of the choice between independent and correlated error models for the observations. In the submitted manuscript, we had included emulator code uncertainty (i.e., error in the emulator relative to MALI) in the error budget. In the analysis below, we show that with emulator code uncertainty included, the choice between independent and correlated observation error models has little effect on learning. Based on this new understanding, we are also making a second methodological change to the study: excluding emulator code uncertainty and assuming that the future trajectories from the **emulator**, rather than MALI, are the zero-error reference future. This allows us to isolate future learning relative to a defined reference without conflating it with uncertainty from the present emulator approximation. Given that the MALI simulations themselves represent model-based future ice-sheet states rather than future reality, and certainly include significant model error, this seems like a beneficial tradeoff.

Thus, in our revised manuscript we will regenerate all results with two major methodological changes: 1) switch from independent to temporally correlated observational uncertainties, 2) switch from including emulator code uncertainty to assuming the emulator trajectories represent the states we are attempting to learn. We have already recalculated all results in the study using these two changes, and all responses throughout the rest of the document are informed by the updated results. The major conclusions about future learning for this problem remain unchanged, but many of the specific details change, and the Results and Discussion sections will be extensively updated to account for this. In responses below, we point out places where we know this will be the case.

In the rest of this section, we provide a detailed explanation of the analysis we did to understand the effects of these two methodological changes (observation correlation and emulator code uncertainty) and justify our decisions for updating our results and analysis.

Observation error covariance. Let r_s denote the annual grounded mass-change rate in year s , with observational error variance σ_a^2 , where $\sigma_a = 5.720 \text{ Gt yr}^{-1}$. Assuming independent observational errors across years, as in the submitted manuscript, the cumulative grounded mass change at analysis year t is

$$C_t = \sum_{s=2017}^t r_s, \quad \text{Var}(C_t) = \sigma_a^2(t - 2016) = \sigma_{o,t}^2,$$

This is the linearly increasing variance already used in the submitted manuscript. However, cumulative observations at different analysis years are correlated because they share the same annual observational errors up to the earlier year. For $t' \leq t$,

$$\text{Cov}(C_t, C_{t'}) = \sigma_a^2(\min(t, t') - 2016), \quad \text{Corr}(C_t, C_{t'}) = \sqrt{\frac{\min(t, t') - 2016}{\max(t, t') - 2016}}.$$

Thus, the observational errors associated with the cumulative observations $\mathbf{C} = (C_{2030}, C_{2045}, \dots, C_{2300})$ follow a multivariate Gaussian distribution with Wiener covariance (Higham, 2001), where $[\Sigma_o]_{t,t'} = \sigma_a^2(\min(t, t') - 2016)$. Going forward, we use this full covariance matrix in place of the diagonal covariance matrix used in the submitted manuscript.

Incorporation of correlated observation errors. We now use the Wiener covariance in both the synthetic future observation generation and the likelihood function. For each synthetic future-observation trajectory, we now draw the observation error jointly as $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \Sigma_o)$ and add it to the emulated cumulative grounded mass change mean. For the sequential calibration at analysis year t , the cumulative observations through t are evaluated jointly using a multivariate Gaussian likelihood with covariance $\Sigma_o^{(t)}$. The additional cost is small because the covariance matrix has at most 19×19 entries, and no changes to the NUTS sampler are required. If emulator code uncertainty is included, it is added to the observation error covariance using an additional diagonal matrix $\text{diag}(\sigma_{c,t',i}^2)$. The three present-day (2015) observables continue to be treated as independent, as in Jantre et al. (2024).

Comparative analysis of error-model configurations. To incorporate the Wiener covariance, we revisited the calibration and uncertainty propagation workflow and came across an unrelated inconsistency in how the GP emulators were queried. Correcting this issue defines **scenario A'**, which otherwise retains the modeling choices of the submitted manuscript. Next, we added the Wiener covariance to this baseline to define **scenario B**. Scenarios A' and B produced nearly identical learning results. We suspected that the emulator code uncertainty, which exceeds the observation error uncertainty for the future cumulative grounded mass-change observations, might be the reason for negligible change in future learning. We therefore also evaluated both observation error models without emulator code uncertainty, defining **scenario C** (independent) and **scenario D** (Wiener). This gives a 2×2 comparison of independent versus Wiener observation errors, with and without emulator code uncertainty (Table 1).

Correction to the GP emulator query. In the synthetic future-observation generation, parameter calibration, and forward uncertainty propagation with emulator code uncertainty, we had used the `predict_y` function from the `GaussianProcesses.jl` Julia package. This function includes a fixed nugget in the GP predictive variance. Because we treat the MALI model as a deterministic simulator, we instead use `predict_f`, which returns the GP predictive variance without this additional observation-noise nugget term. Scenario A' uses `predict_f` while retaining all the modeling choices from the submitted manuscript including independent observation errors and emulator code uncertainty. **Scenario B** differs from A' only by using the Wiener covariance for the cumulative grounded mass-change observations.

Effect of the Wiener covariance with emulator code uncertainty. Scenarios A' and B are nearly indistinguishable. The Wiener covariance changes the learning by less than 2.5% for every parameter and projection horizon, with final-year 90% credible-interval width ratios of 0.98–1.00 (columns A' vs. B of Table 1). The emulator code uncertainty dominates the total error covariance at the analysis years where most learning occurs, making the effect of the observation covariance small in these configurations. Thus,

correcting the temporal covariance alone has little effect on the future-learning results obtained with the emulator code uncertainty present.

Analysis without emulator code uncertainty. Emulator code uncertainty represents approximation error from replacing MALI with a GP emulator and can, in principle, be reduced by increasing the MALI training ensemble. We therefore also recomputed the analysis without this uncertainty in the synthetic future-observation generation, calibration likelihood, and forward uncertainty propagation. These configurations define **scenario C** (independent) and **scenario D** (Wiener). Removing emulator code uncertainty results in substantially greater learning in both configurations (columns C, D of Table 1). We will adopt scenario D as the primary configuration in the revised manuscript because it uses the corrected Wiener observation error model and evaluates future learning without the additional limitation imposed by the present emulator approximation. Scenarios A' and B are retained in this response letter only to show the effect of the covariance correction when the present emulator code uncertainty is included.

Effect of the observation error covariance without emulator code uncertainty. In scenarios C and D, the Wiener model produces final-year parameter credible intervals that are approximately $2\text{--}3\times$ narrower than those of the independent model (columns C vs. D of Table 1). The Wiener likelihood is equivalent to using the independent increments of the cumulative observations between successive analysis years, whereas the independent model uses the cumulative levels. Therefore, the relative information provided by the two error models depends on the sensitivity of the modeled cumulative grounded mass change to the parameters. For the Amery catchment, these sensitivities exhibit substantial changes during the late transient (Fig. 3), when modeled ice-shelf loss accelerates around 2150–2200. This sensitivity structure favors the increment information used by the Wiener model.

To examine this result separately from the full analysis, we developed a linear-Gaussian inverse modeling illustration, provided in the supporting note at the end of this response letter. The illustration is included here for the reviewer and is not intended for the revised paper. It shows that correlated observation errors can produce either greater or smaller parameter learning than independent errors depending on the sensitivity structure, consistent with Tiedeman and Green (2013). Thus, we do not interpret greater learning as a general consequence of including temporal correlation. It results from the sensitivity structure in our analysis.

Scenarios A' and B also produce nearly identical sea-level-contribution learning (Fig. 2). With emulator code uncertainty included, the two observation error models provide nearly identical projection uncertainties. Without the emulator code uncertainty, the narrower parameter posteriors in scenario D result in narrower sea-level contribution intervals than in scenario C. At 2200 and 2300, the final-year widths in scenario D are approximately 50% and 67% of those in scenario C, respectively. The corresponding fold-reductions are approximately $22\times$ vs. $44\times$ at 2200 and $67\times$ vs. $99\times$ at 2300 (Table 1). As the final-year credible-interval widths are very small, these fold-reductions should be interpreted together with the absolute widths. The difference is smaller at 2100, where the scenario D final-year width is $\sim 89\%$ of the scenario C width.

Independently of this Wiener-versus-independent comparison, scenario D shows near-complete learning of the sea-level contribution at the longer projection horizons. The 90% credible-interval widths contract from approximately 64 and 77 mm to about 1.5 and 0.8 mm at 2200 and 2300, respectively (Table 1, Fig. 2). Thus, once the full observation record is assimilated, little projection uncertainty remains at these horizons. This level of learning is not present in the scenarios A' and B where emulator code uncertainty is included and the corresponding reductions are approximately $6\times$ and $8\times$ at 2200 and 2300, respectively. Emulator code uncertainty affects the projection uncertainty through both the calibration likelihood and the predictive uncertainty added during forward uncertainty propagation through the PCA-GP emulator. A separate diagnostic, in which we held the parameter posteriors obtained with emulator code uncertainty (scenarios A' and B) as fixed and removed only the PCA-GP predictive variance, confirmed that the latter contributes to the residual projection uncertainty, particularly at the longer horizons. We omit the diagnostic details here for brevity.

Supporting figures. As we plan to keep the revised manuscript lean, the diagnostics that establish the Wiener-vs-independent mechanism and the effect of including or excluding emulator code uncertainty are collected here in the response letter: the parametric learning curves for all four configurations (Fig. 1), the corresponding sea-level-contribution learning curves (Fig. 2), the emulator sensitivity diagnostics (Fig. 3), and the linear-Gaussian inverse modeling illustration in the supporting note at the end of this letter.

Planned revisions to the manuscript. We will use scenario D, with Wiener observation errors and

without emulator code uncertainty, as the primary configuration in the revised manuscript and update the parameter and sea-level-contribution learning results, figures, and quoted values in Sections 3–4 accordingly.

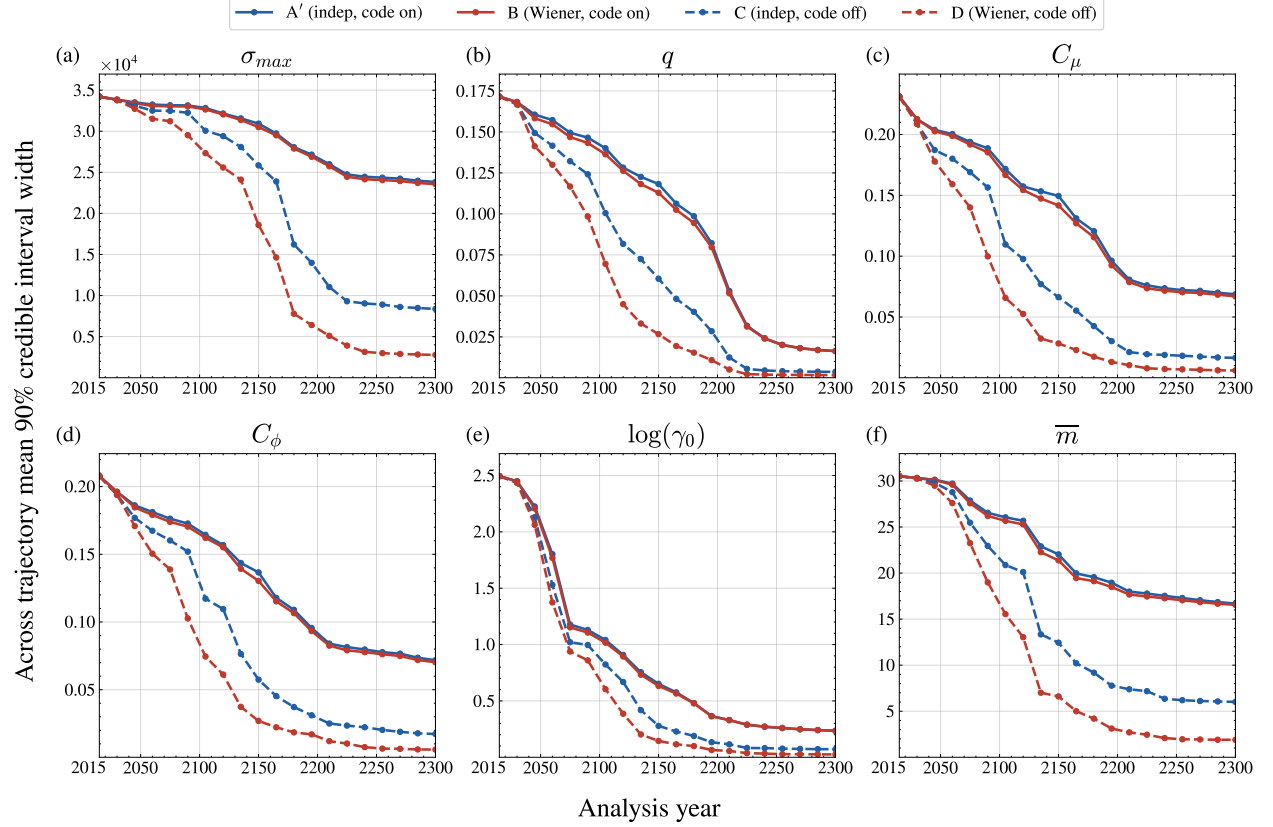


Figure 1: Parametric learning curves for the four error-model configurations: across-trajectory mean of the 90% credible-interval width of each calibrated MALI parameter versus analysis year (2015 is the present-day posterior year). Solid curves include emulator code uncertainty (A' blue, B red), while dashed curves exclude it (C blue, D red). When emulator code uncertainty is included, parameter learning curves are nearly coincident. Without emulator code uncertainty, parameter learning increases substantially, and the Wiener model (D) produces narrower intervals than the independent model (C) for every parameter.

Specific comments

L75: Say if all parameters are scalar.

All six calibrated MALI parameters are scalars, and we will state this explicitly at first mention.

I am also unsure whether there are also parameter fields lurking behind that are not being treated probabilistically. For instance, does MALI carry out an adjoint-based calibration of basal friction coefficients in order to reproduce observed velocity patterns, with C_μ simply a scaling of these values? (And similarly for ice stiffness.) If so, it is worth commenting on whether treating these as “certain” has implications for the study.

The reviewer’s inference is essentially correct. C_μ and C_ϕ are spatially uniform scalar multipliers applied to spatially varying, time-invariant basal-friction (μ) and ice-stiffness (ϕ) fields, respectively. These fields are obtained during model initialization and are held fixed in our study; only their scalar multipliers C_μ and C_ϕ are treated as uncertain. As described in Jantre et al. (2024), the spatial fields were obtained through PDE-constrained optimization using observed surface velocities, following the adjoint-based optimization framework of Perego et al. (2014).

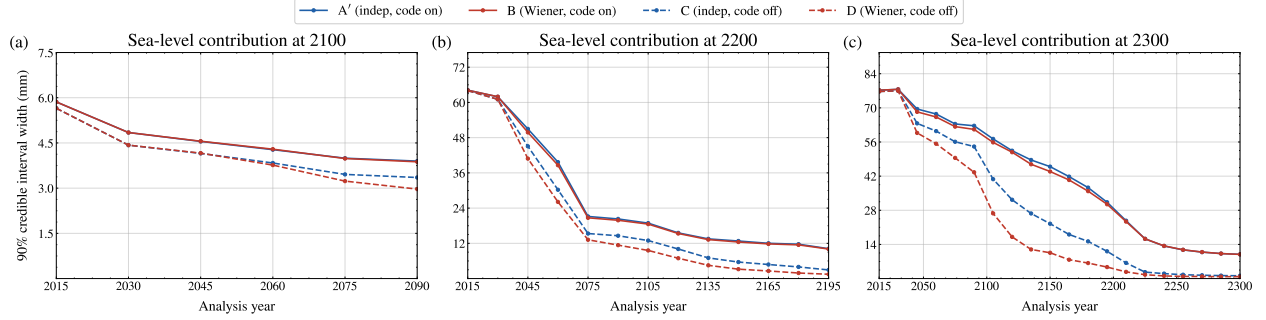


Figure 2: Sea-level-contribution learning curves for the four error-model configurations: across-trajectory mean of the 90% credible-interval width (mm) of the projected sea-level contribution at (a) 2100, (b) 2200, and (c) 2300 versus analysis year. Colors and line styles are as in Fig. 1. Each scenario starts from its own scenario-conditional 2015 baseline, and these baselines are nearly identical across scenarios. With emulator code uncertainty included (A', B), learning is modest, with approximately $6\times$ and $8\times$ reductions at the final analysis year for 2200 and 2300 sea-level projections. Without emulator code uncertainty, learning increases substantially, and the Wiener model (D) produces narrower intervals than the independent model (C) at every horizon (C vs. D about $22\times$ vs. $44\times$ at 2200, and $67\times$ vs. $99\times$ at 2300).

Table 1: Future learning under the four error-model configurations: percentage reduction in the 90% credible-interval width from 2015 to 2300 for each calibrated parameter (across-trajectory mean, top block), and the sea-level-contribution learning factor (fold reduction in width) at each projection horizon (bottom block). A' and B use independent and Wiener observation errors with emulator code uncertainty included, respectively. C and D use independent and Wiener observation errors without emulator code uncertainty, respectively. A' and B produce nearly identical learning. Without emulator code uncertainty, learning is substantially greater, with the Wiener model (D) producing greater learning than the independent model (C) for both the parameters and sea-level contribution. Each column's 2015 sea-level-contribution baseline uses the corresponding emulator code uncertainty setting.

	A' (indep., code incl.)	B (Wiener, code incl.)	C (indep., code excl.)	D (Wiener, code excl.)
σ_{max}	30.4%	31.1%	75.5%	91.9%
q	90.5%	90.5%	97.9%	99.0%
C_μ	70.3%	71.0%	92.9%	97.4%
C_ϕ	65.5%	66.2%	91.7%	97.2%
$\log \gamma_0$	90.6%	90.6%	97.1%	99.0%
$\bar{m}$	45.3%	45.8%	80.3%	93.8%
SLE 2100	1.50 $\times$	1.51 $\times$	1.68 $\times$	1.90 $\times$
SLE 2200	6.30 $\times$	6.38 $\times$	21.8 $\times$	43.6 $\times$
SLE 2300	7.77 $\times$	7.80 $\times$	66.6 $\times$	99.3 $\times$

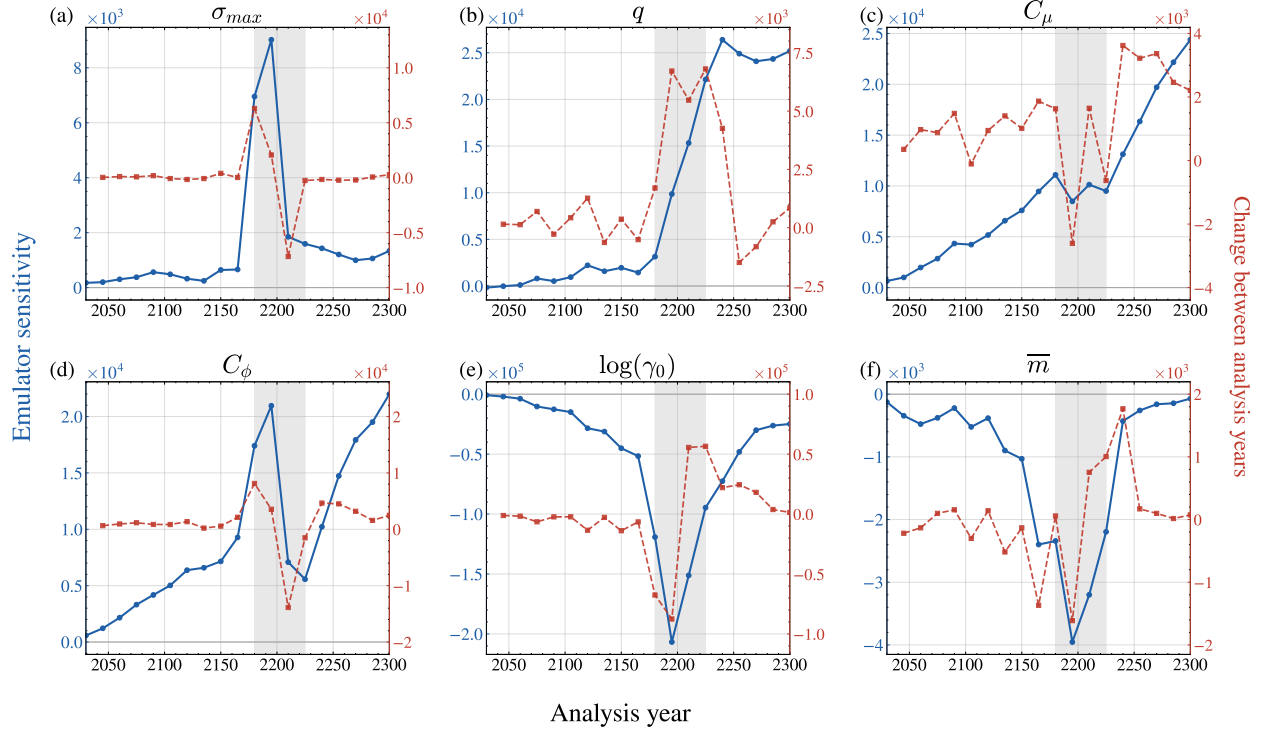


Figure 3: Emulator sensitivity to parameters evaluated at the present-day posterior-mean parameters. Each panel shows the sensitivity of the cumulative grounded mass-change emulator (blue, left axis) and its change between analysis years (red, right axis) versus analysis year. Substantial sensitivity changes are concentrated in a late, sharp transient (shaded band, 2180–2225). Because the Wiener likelihood conditions on the *increments between successive analysis years* of the cumulative observations, this sensitivity structure can make the correlated model more informative than the independent model.

Treating these spatial fields as fixed limits the scope of our uncertainty analysis. The scalar multipliers represent uncertainty in the overall magnitude of basal friction and ice stiffness while preserving their inferred spatial patterns. We therefore do not account for spatially heterogeneous uncertainty in these fields. Such uncertainty could affect the modeled dynamical response, but we cannot determine from the present analysis how including uncertainty in these spatial fields would change the amount of future parameter or projection learning, and therefore do not speculate on its direction or magnitude. Our reported learning should instead be interpreted as conditional on the initialized spatial fields. We will emphasize this limitation in our revisions.

L82: there is mention of a GP emulator; readers should not need to refer to another paper to see this. Could a brief intuitive explanation be given? My understanding from Wernecke 2020, whom you cite, is that an PCA (SVD) analysis is done and the temporal expressions of the spatial eigenvectors are the things that are replaced by GPs (which are, as I understand it, essentially brownian bridges). A truncated set of spatial eigenvectors is retained; can you say how many are retained, and how this decision impacts posterior uncertainty? And that said, I might be misunderstanding, so an intuitive explanation would be helpful.

We will add a brief explanation of the Gaussian process (GP) emulator at first mention. A GP emulator is a statistical surrogate model that learns the relationship between the six MALI parameters and an expensive MALI simulation output, allowing for rapid predictions at new parameter values within the interpolation regime. For the VAF-change time series used to calculate sea-level contribution, we apply PCA to obtain a reduced temporal representation of the ensemble, rather than spatial eigenvectors. We then emulate each retained principal-component score as a function of the six MALI parameters using a separate GP and reconstruct the VAF-change time series predictions from these emulated scores. We retain the five leading

principal components, which together explain more than 99% of the variance in the VAF-change time-series outputs from the MALI ensemble (Jantre et al., 2024). The remaining components are omitted from the reconstruction, so the variability represented by them is not included in the predictive distribution. We have not separately quantified the effect of this truncation on the final posterior predictive uncertainty. We will also clarify that the GP emulators use a smooth Matérn-2.5 covariance kernel and are not Brownian bridges.

L93: need to say how this is defined (is grounded mass volume above floatation), how error is defined, either here or point to the section where it is done. We should not need to consult jantre to get the basic idea.

We will define grounded mass change at first mention. It is the change in the total mass of grounded ice in the Amery catchment (ice resting on the bed) and is distinct from volume above flotation (VAF) change, which is used to calculate sea-level contribution. The observational uncertainty is based on the grounded-ice mass-balance estimate of Jantre et al. (2024), with an annual-rate standard deviation of $\sigma_a = 5.720 \text{ Gt yr}^{-1}$. For the cumulative future observations, this annual uncertainty gives the time-dependent Wiener covariance described in General comment 2.

L114: present day posterior: where do you discuss the present day posterior density calculation.

The present-day (year 2015) posterior is the calibrated posterior from Jantre et al. (2024), obtained by conditioning the six MALI parameters on the three present-day observational constraints (mass balance, grounding-line movement, and calving-front movement). We use this posterior as the starting point for the future-learning analysis and will clarify this at first mention in Section 2.

L125: I think the specific form(s) of Bayes' rule need to be given here. it is difficult to envision the likelihood function. And while it becomes clear later, it is not clear here that you are treating each of the 100 trajectories as a separate realization to be learned from in order to find statistics of learning, that should be made clear.

We will make the likelihood function explicit. For a single synthetic future-observation trajectory, the posterior at analysis year t is

$$p(\boldsymbol{\theta} \mid \mathcal{D}_t) \propto p(\mathbf{y}_{2015} \mid \boldsymbol{\theta}) p(\mathbf{C}_{\leq t} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

where $\mathbf{y}_{2015}$ denotes the three present-day observables, $\mathbf{C}_{\leq t} = (C_{2030}, \dots, C_t)$ denotes the cumulative future observations available through year t , and $\mathcal{D}_t = (\mathbf{y}_{2015}, \mathbf{C}_{\leq t})$. In the revised analysis, the likelihood for the cumulative future observations is

$$p(\mathbf{C}_{\leq t} \mid \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{C}_{\leq t}; \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}_o^{(t)}\right),$$

as described in General comment 2. We will also make explicit that calibration is performed separately for each of the 100 synthetic trajectories at each analysis year, and the resulting learning is then summarized across the 100 trajectories.

L137. This seems like a zero-order error you are introducing, and I do not see why you would do it. If the loss from, say, 2100 to 2110 is independent of the loss from 2015 to 2100, and both are gaussian, then the loss from 2015 to 2110 is the sum of the two, and finding the statistical properties of the correlation between this and loss to 2100 seems a simple known problem. By treating cumulative loss to 2100 and cumulative loss to 2110 as independent you are surely skewing the posterior uncertainties in a nonnegligible way — ignoring the correlations could overestimate reduction in uncertainty.

We agree that this is a specific example of General comment 2. As shown there, $\text{Cov}(C_t, C_{t'}) = \sigma_a^2(\min(t, t') - 2016)$, and we now account for this Wiener covariance in both the synthetic future observation generation and the likelihood. Please see our response to General comment 2 for the formulation and resulting re-analysis.

L144: this makes more sense now but I hadn't realized you are treating each observational trajectory independently.

Correct. As noted in our response to L125, we will make explicit in Section 2 that the 100 synthetic future-observation trajectories are independent and that the parameters are calibrated separately for each trajectory at each analysis year.

L148-151: confused — this is how you generated the synthetic data, no? Why are you introducing it again?

L152-153: again confused – this seems a necessary step to (a) synthetic observations and (b) likelihood evaluation. yet it is badged as "probabilistic future learning". or maybe this is the forward propagation of uncertainty. as it is not clear how this is done (monte carlo or MCMC?) please give more detail on this, using eqns if necessary.

L154-156: isn't this what you said you did at lines 144-146? I'm very confused now.

We apologize for the confusing organization. We will restructure Section 2 to distinguish three separate steps:

- *Generation*: create the 100 synthetic future-observation trajectories from parameter draws from the present-day posterior;
- *Inference*: for each trajectory and analysis year, recalibrate the parameters using MCMC (NUTS);
- *Propagation*: propagate samples from each recalibrated parameter posterior through the PCA-GP emulator of VAF change to obtain sea-level-contribution distributions at 2100, 2200, and 2300.

The last step is forward uncertainty propagation using Monte Carlo samples, not an additional MCMC calibration. We propagate 5000 samples from each recalibrated posterior through the PCA-GP emulator and use the resulting sea-level-contribution samples to quantify projection uncertainty.

Figure 2, "Sample from present-day posterior" in legend. I am not sure I understand this — are these the parameters in the specific realization in the 100-member ensemble used for synthetic generation?

Yes. The black dash-dotted line marks the parameter value drawn from the present-day posterior and used to generate the synthetic future-observation trajectory under consideration, as stated in the figure caption.

L217: 4.17 [2.80–5.09] mm. My understanding of what this describes is WAY off, which needs to be remedied, or this does not make sense. It is saying that with the information available in 2090, there is *uncertainty* of about 4 mm SLR in the cumulative SLR by 2100. I don't see how this is possible, unless all of the SLR from 2015-2100 is presumed to happen in the last decade. More generally, and at least assuming Gaussian statistics, I would expect the uncertainty in a future value of a cumulative sum at a certain time to go to zero as that time is approached. And this is precisely because the cumulative sums at different times correlate with each other. And so I wonder if what I'm seeing is a consequence of treating these cumulative sums as independent.

The quantity observed for calibration and the quantity being projected are different. The synthetic future observations are of cumulative *grounded mass change*, whereas the 2100 sea-level contribution is obtained from VAF change by propagating recalibrated MALI parameter samples through the PCA-GP emulator for VAF change. Observations through 2090 therefore constrain the parameter posterior and do not directly observe the cumulative sea-level contribution itself. Consequently, there is no requirement that the uncertainty in the 2100 sea-level contribution vanish as the observation record approaches 2100. The quoted 4.17 mm in the submitted manuscript is the *width* of the 90% credible interval, not 4.17 mm of sea-level contribution occurring between 2090 and 2100. After correcting the temporal covariance as described in General comment 2, the 2100 credible-interval width remains nonzero, showing that the nonzero width is not solely a consequence of treating the cumulative grounded-mass observations as independent.

L239-241. interesting. as the level of buttressing is not always the same, is the degree of buttressing a factor as well, and can this be looked at in this context?

Yes. The degree of buttressing, rather than simply the presence or area of floating ice, can affect how strongly ice-shelf change alters the grounded-ice stress balance and therefore the dynamical response that informs q . Our present analysis does not explicitly quantify buttressing, and there is also not a single agreed-upon definition or metric for buttressing. We therefore do not attempt to relate the learning rate to a specific buttressing metric. Instead, we interpret the rapid learning about q more generally as coinciding with the strong dynamical response to substantial ice-shelf loss. Note also that this discussion will change with the recalculated learning rates that account for covariance and ignore emulator code uncertainty. This question of changing buttressing will be acknowledged in the rewritten version.

L243-244: I disagree with interpretation. if floating ice is not providing backstress its removal will not affect discharge, providing little information/learning even if ice is lost rapidly.

We agree with the reviewer that rapid loss of floating ice is not by itself sufficient to produce informative learning. The relevant mechanism is the resulting change in grounded-ice dynamics. We will reword this passage to attribute the learning about q to the strong dynamical response during this period, rather than to rapid ice loss itself. Like the previous response, the discussion here will be adjusted for the recalculated learning rates, and we will address this buttress questioning, should it remain relevant, as part of those revisions.

L279, point 2: this is true of the basal friction coefficient scaling too, comment?

Agreed. The revised analysis substantially changes the learning behavior of C_μ (see General comment 2), so the discussion at L279 will be revised accordingly and the original interpretation will no longer apply.

L290–291: I think it would be useful to do a Fürst-type analysis of the buttressing number as his numbers do not apply to these new modeled configurations of the ice shelf.

We agree that a buttressing diagnostic could provide a more direct measure of the mechanical support provided by the evolving ice shelf. We do not currently quantify buttressing explicitly, so we do not use a buttressing number to interpret the learning results. We instead interpret the enhanced learning of q more generally in terms of the strong dynamical response during the period of substantial ice-shelf loss. We prefer not to conduct formal buttressing analysis, in part because it becomes tricky to decide which full MALI simulation to evaluate and when.

L296–297: again, this does not make sense to me — such uncertainty for something 10 years in the future? means im missing something in my understanding

This is the same issue raised at L217. The observations through 2090 are of cumulative grounded mass change and constrain the MALI parameter posterior, whereas the 2100 sea-level contribution is obtained from VAF change by propagating the recalibrated parameter samples through the PCA–GP emulator for VAF change. The uncertainty in the 2100 sea-level contribution therefore need not vanish as the observation record approaches 2100. The 2100 projection width also remains nonzero in the revised analysis with the Wiener covariance (General comment 2).

L304–306: is it just the change in forcing, or does the change in the system itself play a role? Thwaites glacier is likely to keep retreating and thinning greatly over the next 50–100 years without any changes to forcing.

Both can matter. In our simulations, faster learning coincides with a period of changing prescribed forcing and strong dynamical response of the ice sheet, particularly during substantial ice-shelf thinning and loss. We do not separately isolate the contributions of forcing changes and internally evolving ice-sheet dynamics, so we do not attribute the learning to either one alone. The reviewer’s example of Thwaites highlights that substantial dynamical change could also provide a strong learning signal even under weakly changing forcing. We will revise the discussion to distinguish the roles of changing external forcing and the evolving dynamical response of the ice sheet. We will also note that while previous studies have identified Thwaites Glacier as particularly susceptible to internal dynamics, the Amery Ice Shelf catchment is believed to be less susceptible.

L313–314: is this because the projections at 2100 are so small that, in absolute terms, they are still well below uncertainties at 2200?

Yes. The projected sea-level contribution at 2100 is small in absolute magnitude, and its credible-interval width is much smaller than at 2200 and 2300. This is distinct from the amount of *learning*, which we measure by the contraction of the credible-interval width as observations accumulate.

L331, “slightly larger than the mid-trajectory mean”, please remind us what parts of the figure to look like. I thought the uncertainty of the credible interval was the beige shaded area. the half-width of this is far less than the across trajectory mean.

The shaded band represents the across-trajectory 66% range of the 90% credible-interval widths, and the thick curve represents their across-trajectory mean. This discussion will be revised together with the updated sea-level-contribution learning results described in General comment 2.

L339: remove “an”. “more gradual” than what??

We will remove the stray “an”. The corresponding discussion of the relative learning rates at 2200 and 2300 will also be revised with the updated sea-level-contribution results described in General comment 2.

L346: this is interesting but please show rigorous statistics on these claims of relationships in Figs 9 and 10.

We quantified the relationships discussed in Figs. 9 and 10 from the original manuscript using Spearman rank correlations between the parameter value used to generate each of the 100 synthetic trajectories and the corresponding 90% credible-interval width. These relationships persist in the revised analysis (General comment 2). For the 2200 sea-level contribution (calibration through 2075), the relationship with $\log(\gamma_0)$ is strong ($\rho = 0.762$, $p = 3.2 \times 10^{-20}$). For the 2300 sea-level contribution (calibration through 2150), the relationship with q is also strong ($\rho = -0.655$, $p = 1.4 \times 10^{-13}$). We will report these statistics along with their relationships in the revised manuscript. The corresponding revised sensitivity plots (Figs. 4 and 5) are shown below for completeness.

L416–417: A very good point.

We thank the reviewer for highlighting this point.

L429: the answer then is to not use diagonal covariance matrices, even if it takes longer to converge.

Agreed. As described in General comment 2, we will replace the diagonal covariance for the cumulative grounded mass-change observations with the full Wiener covariance Σ_o . We will update the Limitations text accordingly so that temporal correlation is no longer presented as an unaddressed limitation.

L431: “in reality” — please explain why this is so difficult to incorporate?

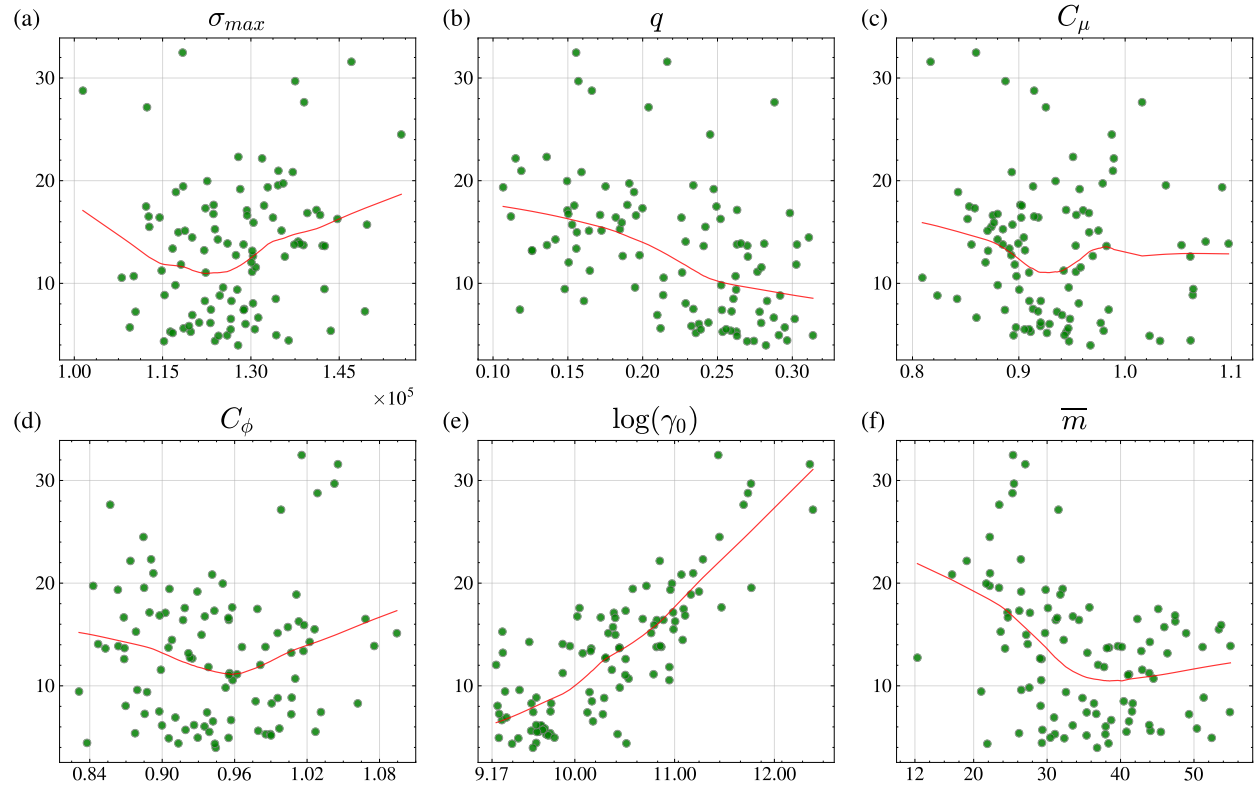


Figure 4: **Revised sensitivity of future learning of year-2200 sea-level contribution to individual MALI parameters informed by observations through 2075 (scenario D).** Points show, across the 100 synthetic future-observation trajectories, the year-2200 sea-level contribution 90% credible-interval width versus the parameter value used to generate each trajectory. Red curves show LOWESS trends.

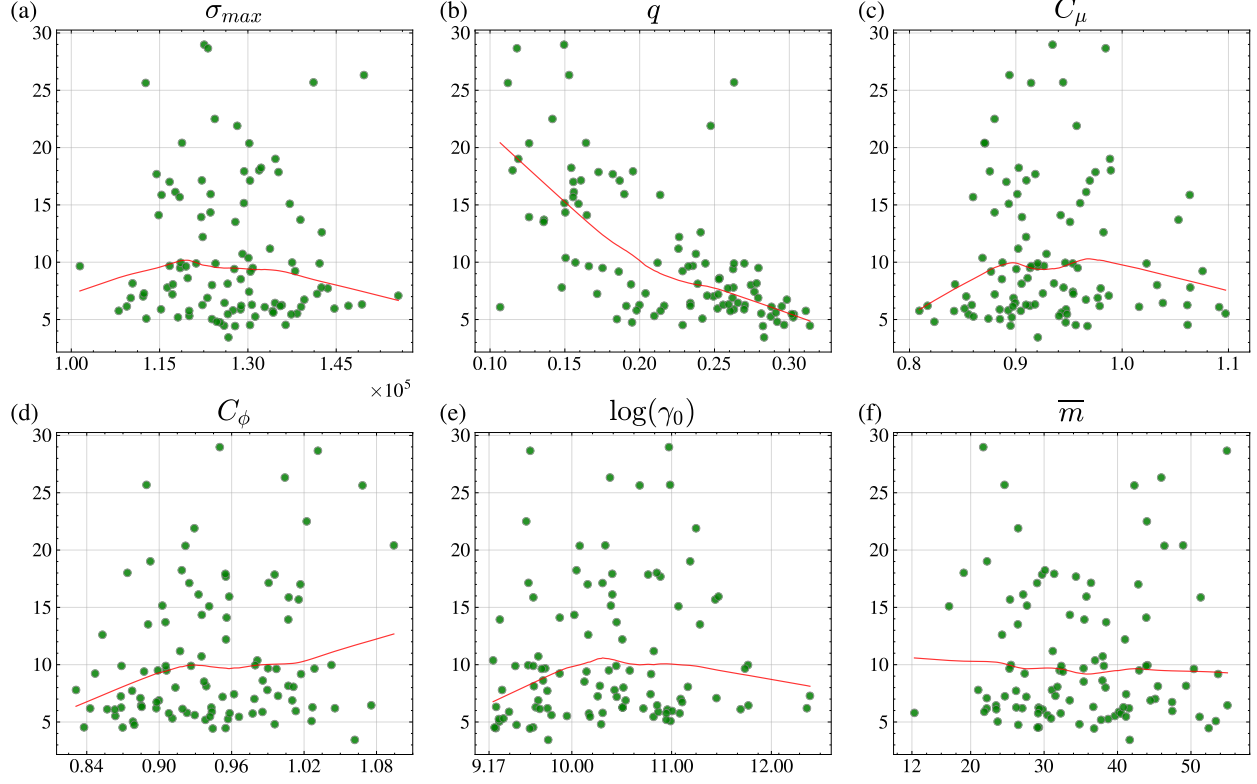


Figure 5: **Revised sensitivity of future learning of year-2300 sea-level contribution to individual MALI parameters informed by observations through 2150 (scenario D).** Points show, across the 100 synthetic future-observation trajectories, the year-2300 sea-level contribution 90% credible-interval width versus the parameter value used to generate each trajectory. Red curves show LOWESS trends.

Temporal correlation of cumulative grounded mass change is straightforward to incorporate and will be treated with the Wiener covariance (General comment 2). The greater difficulty is in specifying and estimating appropriate covariance structures among different observables and for spatially resolved observations, where those correlation structures are not readily available and would need to be modeled or estimated. We will add this distinction in the revised Limitations section.

L440: Recinos reference: here, overconfidence was due to unrealistically small errors in observations. (Which I think you are admitting by treating cumulative mass losses as independent..)

The temporal-independence assumption for cumulative grounded mass-change observations mis-specifies the observation-error covariance and will be corrected as described in General comment 2. Separately, our perfect-model assumption omits model structural error, which can lead to overconfident inference. For this point we cite Recinos et al. (2023). We will revise the text to distinguish these two issues clearly.

Section 4.5: I think this would benefit from a review of the literature around Optimal Experimental Design.

We will expand Section 4.5 to better place our future-learning framework within the Bayesian optimal experimental design literature. In addition to Rainforth et al. (2024), which is already cited, we will include relevant work by Chaloner and Verdinelli (1995), Ryan et al. (2016), and Huan et al. (2024).

Response to Reviewer 2 (Eliot Jager)

Review summary. *This article examines the potential impact of future observations on reducing uncertainty in projections for the Amery Ice Shelf catchment, using a Bayesian framework. It uses an emulator of the MALI ice-sheet model with six input parameters that modulate melting, friction, rheology, and calving, and produces sea-level rise as output. The study then uses present-day cumulative ice-mass change in the catchment to refine the projections, before examining how new idealized future observations might further reduce uncertainty in future projections. It shows that substantial reductions in both SLR projection uncertainty and parameter uncertainty occur when major changes in ice dynamics take place.*

The idea is highly original and, to my knowledge, is applied here for the first time in an ice-sheet modeling context. Moreover, it raises an interesting question that is also policy-relevant. The introduction, results, and discussion are well written, and the overall structure is easy to follow. However, I have some general comments, mainly concerning the methodology.

Major comments

Major comment 1. *I am not a specialist in GP emulators, although I understand the general idea behind them. The methods section was a bit hard to follow, especially for a non-specialist like me, and would benefit from clarification to make it accessible to a broader audience. Of course, I understand that you do not want to repeat all the details from Jantre et al. (2024), and I am not asking you to add more technical detail, but rather to be clearer, to properly define the terms you introduce, and to remain consistent throughout the text (including in the Results and Discussion sections). For example, in my minor comments, I refer to terms such as “best estimate” and “trajectory”, the latter of which seems to be used in different contexts. I would also suggest adding a figure to illustrate the framework, which might help non-specialists understand and follow the methods section; see, for example, Fig. 2 in Hill et al. (2021). The rest of the manuscript, including the results and discussion, is easy to follow for non-specialists, so it is only the methods section that needs clarification to make the article understandable to all ice-sheet modelers, if that is your goal.*

We will revise Section 2 to improve its accessibility and define recurring terminology clearly at first mention. As described in our response to Reviewer 1’s General comment 1, we will begin Section 2 with an overview of the full workflow before describing each step in detail. We will also clarify the meaning of “trajectory” and use the term consistently throughout the manuscript to refer to one synthetic future-observation time series generated from a single parameter vector drawn from the present-day posterior. We will not add a separate schematic, but will make the workflow explicit in the revised text.

Major comment 2. *I also have some concerns about C_ϕ C_μ , which also apply to Jantre et al. (2024). I am not sure I understand their purpose or physical meaning. My understanding from Jantre et al. (2024) is that the ϕ and μ fields, representing rheology and friction, are first calibrated to match an observed velocity field. These calibrated fields are then multiplied by two spatially uniform parameters, C_ϕ and C_μ . However, doing so might produce velocity fields that differ substantially from observations. This may help fit mass changes when the C values differ from 1, but potentially for the wrong reasons, for example if mass loss occurs in the wrong locations compared with observed dh/dt fields. For instance, if the model does not produce enough mass loss, reducing friction everywhere could lead to mass loss in the interior, where it should not occur. Here are some examples that take into account uncertainty from calibration, which makes more physical sense to me: some from my own articles, Jager et al. (2024a,b), as well as other approaches such as Recinos et al. (2023) and Jakeman et al. (2025). I understand that this is not the core of the present paper, and that it is more something that should perhaps have been addressed in Jantre et al. (2024), where it seems to be mentioned only briefly in the limitations: “In particular, the ice stiffness and basal friction scaling factors are a highly simplified means of representing uncertainties in what are actually two-dimensional spatial fields.” I am not sure what the best way is for the authors to address this comment, because removing this approach in order to better match current observations would require a lot of work. However, since this is not the core of the present paper, I see one possible way to improve or clarify this point, though it is not mandatory and the authors may prefer another approach. The authors could mention and discuss that these parameters do not have a specific physical meaning, but instead allow them to create an ensemble of synthetic, artificial,*

and non-physical friction and rheology fields. In this sense, they do not represent reality directly, but rather different possible alternative model realities.

C_μ and C_ϕ are spatially uniform scaling factors applied to the spatially varying basal-friction and ice-stiffness fields, respectively. As described in Jantre et al. (2024), the underlying fields are obtained through PDE-constrained optimization using observed surface velocities. The ranges considered for C_μ and C_ϕ were then selected from sensitivity tests based on observational uncertainty and variability in the resulting velocity fields. Thus, these parameters represent uncertainty in the overall magnitude of the initialized basal-friction and ice-stiffness fields while preserving their inferred spatial patterns.

We agree that this does not represent uncertainty in the spatial patterns of the two fields. As also noted in Jantre et al. (2024), this is a simplified representation of uncertainty in quantities that are spatially varying fields. Our analysis is therefore conditional on the inferred spatial patterns of μ and ϕ . We will emphasize this interpretation when introducing C_μ and C_ϕ and in the limitations. We chose this approach in part because the future learning methodology is complicated enough to conduct and to communicate without introducing substantial additional complexity with high-dimensional spatial uncertainties.

Major comment 3. *A point related to the previous comment is that I am wondering whether it always makes sense to calibrate all the parameters considered here. For example, the average basal melt rate may change over time, so calibrating it to present-day conditions might be reasonable for the near future, but it would likely be different after 100 years. For q , I am quite confident that it can be kept fixed, whereas for other parameters, such as σ_{max} and $\log(\gamma_0)$, I do not know whether it makes sense to keep them constant, as this is beyond my expertise. For C_ϕ and C_μ , keeping them fixed might make sense if the interpretation suggested above is adopted, as the aim would then be to identify the most appropriate global friction and rheology fields.*

We treat the six uncertain MALI parameters as time invariant over each simulation, consistent with the model formulation used in Jantre et al. (2024). We agree that some effective parameter values could potentially change over multi-century simulations, and will clarify this assumption in the revised manuscript. We note, however, that $\bar{m}$ is the historical ice-shelf-averaged basal melt rate used, together with γ_0 , to determine the bias correction δT . It is not a prescribed basal melt rate that remains constant through the simulation. Future basal melt evolves with the changing ocean thermal forcing through the Jourdain et al. (2020) melt parameterization.

Major comment 4. *I finally have a general comment concerning the choice of observation, although this is less of an issue since the manuscript already includes a paragraph on this point in the limitations section. Using cumulative mass loss is clearly not the choice I would have made, especially for a study on the learning rate, because it is an observation that responds slowly and it takes time for a signal to emerge. Today, we already have observations of changes in velocity fields, grounding-line position, and elevation, which seem more appropriate for calibrating some parameters. For example, for friction and rheology parameters, evolving velocity fields would directly inform them and could possibly help resolve the equifinality problem they face (Arthern and Gudmundsson (2010)). Changes in these 2-D spatial fields would provide much more information, and it is not clear that cumulative mass loss can do so: multiple combinations of parameters might lead to the same global value but with different velocity and thickness fields; see Jager et al. (2024b) for a discussion of this global-target problem. It will also take more time for the signal to emerge: first, there are changes in geometry and velocity fields; at the same time, ice discharge changes; and only later do differences in cumulative mass loss between two ensemble members become visible. This leads to relatively late learning, after the massive mass loss of the catchment, which is likely too late for policymakers. I strongly recommend incorporating the points mentioned above into the discussion (the slow response of cumulative mass loss, the potential added value of velocity, grounding-line, and elevation observations, the risk of equifinality when using a global target, and the implications of delayed learning for policy relevance), as they would significantly enhance the manuscript. Furthermore, I believe it would be very beneficial to include a sensitivity test examining how the results vary with alternative targets, such as ice discharge or mass-change rate; this would not, in my view, require too much additional work, and while it is not strictly necessary, adding it in a supplement or appendix would be an excellent addition.*

The limitations associated with using cumulative grounded mass change, including the loss of spatial in-

formation and the potential value of velocity, elevation, and grounding-line observations, are discussed in Sections 4.4–4.5 of the submitted manuscript. We will add the reviewer’s point that these observations may provide earlier indications of dynamical change, which is relevant to the timing of learning. Our decision to use this one integrated observational metric was in part to keep the future learning methodology easier to implement and easier to follow. Producing a credible emulator of spatial fields like velocity is a substantially more difficult task than for a scalar, smoothly evolving metric like cumulative grounded mass change.

We do not plan to add the suggested sensitivity test. As described in Section 2.1, we use cumulative grounded mass change because the modeled rate variables fluctuate substantially, making the cumulative quantity a more stable target for statistical emulation. Comparing alternative targets would require additional emulators and observation-error models, followed by the full calibration and uncertainty propagation analysis for each target. We therefore keep the present analysis focused on cumulative grounded mass change.

Minor comments

L23: “While our perfect-model assumption and simplified likelihood structure...” → I suggest also mentioning the choice of observations here. In addition, I am not sure what you mean by “simplified likelihood structure”.

We will mention in the Abstract that the analysis uses a single future observable, cumulative grounded mass change. Furthermore, “simplified likelihood structure” refers to treating the three present-day observables as independent of one another and independent of the future cumulative grounded mass-change observations, as described in Section 2.2 of the submitted manuscript. Temporal correlation among the cumulative grounded mass-change observations will be accounted for using the Wiener covariance described in our response to Reviewer 1’s General comment 2.

L52–53: I do not see why “little historical change” is an advantage here. With past dynamical changes, you would have been able to use real observations to study the learning rate, for example by starting in the past and adding observations one by one to examine the reduction in uncertainty.

We agree that “little historical change” is not itself an advantage for studying learning. In this study, the limited historical change means that present-day observations provide relatively weak constraints on future dynamical change, leaving greater scope for future learning. We will revise the wording to make this distinction clear. The other point we will make more explicit here is that calibration is simpler when the system can be assumed to be close to steady state, removing potential time-dependence of parameters, but as the reviewer points out, this also provides a weaker constraint. A sector that has already undergone substantial dynamical change could instead allow learning to be studied retrospectively using historical observations, as the reviewer suggests.

L57–58: This is fine, but it may be worth discussing at some point whether the conclusions/results would be different for other sectors where dynamical changes have already been observed in the past, such as the Amundsen sector.

We agree. As noted in the Conclusions, regions already undergoing substantial dynamical change could provide earlier information about the evolving system than the relatively stable Amery sector. We will expand this discussion to note that past dynamical changes also provide historical observations that could be used to constrain model parameters.

L75–77: It would be useful to add an appendix summarizing the model configuration from Jantre et al. (2024), including the definition, physical meaning, and effect on the model of each parameter. This would be especially useful for Sect. 4.1. It may also be helpful to describe the physical meaning of each parameter for non-ice-sheet modelers.

We will add a concise summary of the six calibrated MALI parameters, including their definitions, prior ranges, physical roles, and effects in the model. We will also clarify the interpretation of C_μ and C_ϕ as described in our response to Reviewer 1 at L75.

L83: “their SSP5 projection” → What does “their” refer to?

“Their” refers to Jantre et al. (2024). We will replace “their SSP5 projection” with “the SSP5 projection ensemble”.

L92–93: In Jantre et al. (2024), I think there is also an initial calibration of the model using velocity fields to calibrate ϕ and μ . If you add an appendix on the model configuration, this first calibration should also be mentioned.

The spatially varying μ and ϕ fields are obtained during model initialization through PDE-constrained optimization using observed surface velocities (Perego et al., 2014; Mouginot et al., 2017). This occurs before and separately from the Bayesian calibration of the six scalar parameters considered in this study. We will include this in the model configuration summary described in our response to L75–77.

L95–97: To improve the stability of the rates, would it not have been possible to use a 15-year average mass-loss rate? I do not know enough about GP emulators, but does this not point to a limitation if they cannot use instantaneous rates?

A 15-year-average mass-change rate could provide a smoother target than an instantaneous rate and is a reasonable alternative. We use cumulative grounded mass change because the modeled rate variables fluctuate substantially and the cumulative quantity provides a more stable target for the GP emulators. This is a practical limitation of the present emulation setup rather than a restriction of GP emulation itself.

L106: This part is a bit unclear to me. Do you have one emulator for each 15-year interval, and did you use the original calibration with present-day observations to obtain them?

We use 19 separate scalar GP emulators, one for each analysis year (2030, 2045, ..., 2300). Each emulator maps the six MALI parameters to cumulative grounded mass change at that year and is trained on the SSP5 projection ensemble of Jantre et al. (2024). The present-day posterior is not used to train these emulators. It is subsequently used to draw the 100 parameter vectors for generating the synthetic future-observation trajectories.

L109–110: “Emulator skill is assessed via five-fold cross-validation (as in Jantre et al. (2024)); after validation, the emulators are refit on the full SSP5 ensemble for subsequent use.” → Similarly, it is unclear to me what exactly is being done here.

We will make this description clear. In five-fold cross-validation, each fold is held out once and predicted using an emulator trained on the remaining four folds. The predictions from the five held-out folds give an out-of-sample prediction for every simulation in the ensemble, which we use to assess emulator skill through the validation diagnostics in Appendix A. After validation, each emulator is refit using the full SSP5 ensemble for subsequent use.

L104–113: The placement of this paragraph on the emulator is a bit surprising. Is the emulator not common to both observation generation and the rest of the analysis? If so, why is this paragraph placed here? Should it not be in a separate subsection on the emulator, or in the introductory paragraphs of Sect. 2?

The GP emulators are used in multiple steps. As part of the Section 2 reorganization described in Major comment 1, we will move the emulator description earlier in Section 2 before the steps that use them.

L123–124: “because we cannot know the true future trajectory, we consider the 100 sampled future trajectories based on the Bayesian calibration of model parameters to the present-day observables” → I do not understand this sentence.

We will rewrite this sentence. By “true future trajectory,” we mean the actual future evolution of the real-world Amery catchment, which is unknown. Hence, we consider 100 synthetic future-observation trajectories, each generated using a parameter vector drawn from the present-day posterior, and evaluate future learning across them.

Sect. 2.1: After reading this paragraph, I am not fully sure how the observations are obtained. It sounds as if, every 15 years, an observation is drawn from all possible trajectories, for example a dark-blue point in 2195 and a cyan point in 2210. I do not think this is what you did, so please clarify. If this is what you did, it seems quite unusual.

This is not how the observations are generated. Each of the 100 trajectories is based on one parameter vector θ_i drawn from the present-day posterior. For each θ_i , we obtain the emulator means at all 19 analysis years and draw the corresponding synthetic future-observation trajectory jointly from a multivariate normal

distribution using the Wiener covariance, as described in our revised analysis and response to Reviewer 1’s General comment 2. Observations from different trajectories are not mixed.

$\sigma_{c,t,i}$: *I am not sure how this quantity is obtained, or what its value is.*

In the submitted manuscript, $\sigma_{c,t,i}^2$ is the predictive variance of the year- t scalar GP emulator evaluated at parameter vector θ_i . It represents emulator code uncertainty and was included in synthetic future-observation generation, the calibration likelihood, and forward uncertainty propagation. As described in our response to Reviewer 1’s General comment 2, emulator code uncertainty is excluded from the revised analysis. Consequently, $\sigma_{c,t,i}^2$ will no longer appear in this part of the revised Methods.

$\sigma_{o,t}$: *This linear growth does not make sense to me. If you look at IMBIE3, for example Ootosaka et al. (2023), the total/cumulative mass change uncertainty does not grow linearly.*

In the submitted manuscript, we assume linear growth of the cumulative mass-change *variance*, not its standard deviation. Specifically, $\sigma_{o,t}^2 = \sigma_a^2(t - 2016)$, so $\sigma_{o,t}$ grows with the square root of elapsed time. This follows from treating the annual mass-change errors as independent increments. Ootosaka et al. (2023) use a related approach in their Eq. (7), where cumulative mass-change uncertainty (standard deviation) is obtained from the root sum square of the monthly mass-change-rate uncertainties. Their cumulative standard deviation does not follow an exact square-root dependence because the underlying rate uncertainties vary with time. As described in our response to Reviewer 1’s General comment 2, we account for the temporal correlation between cumulative observations using a Wiener covariance. This changes the covariance between analysis years but retains the same marginal variance $\sigma_{o,t}^2 = \sigma_a^2(t - 2016)$.

L136–138: *“we assume independence among the 2015 observables”* → *This is clearly a strong assumption. For a mass-change rate, I would agree that correlations between two observations can be neglected, since the signal may be somewhat chaotic, but for cumulative mass change the observations are clearly not independent.*

We agree and have addressed this issue in the revised analysis. Please see our response to Reviewer 1’s General comment 2 for details.

L142–146: *This part is quite unclear to me; perhaps a schematic would help. For each sequence of observables, do you obtain 100 sampled parameter vectors and projections?*

For each of the 100 synthetic future-observation trajectories, we perform a separate calibration at each analysis year to obtain a recalibrated parameter posterior. We then draw 5000 samples from each recalibrated posterior and propagate them through the PCA–GP emulator for volume-above-flotation (VAF) change to obtain the corresponding distribution of sea-level contribution projections. We will make this explanation more explicit in the revised manuscript.

L171–172: *“plum-colored trajectory”* → *In Fig. 3, it would be nice to make this more distinguishable, perhaps by using a star instead of a diamond.*

In the revised version of Fig. 3, we mark the plum-colored trajectory with a star instead of a diamond to distinguish it from the other three trajectories. The credible intervals are more clearly separated in the revised analysis. We therefore retain the shaded bands without additional hatching or boundary lines. The revised figure that we will use is shown here in Fig. 6.

Results: It would be useful to easily identify the Fig. 2 trajectory, as well as the four others shown in Fig. 3, across the different result plots. One suggestion would be to give them names, such as trajectory 1, trajectory 2, etc., and use these names in the legends. You could show these trajectories in Fig. 1, and then show only these four trajectories in Figs. 4 and 6.

We do not intend the trajectories shown in Figs. 2 and 3 in the submitted manuscript to be tracked across the subsequent results. Fig. 2 illustrates sequential calibration for one synthetic future-observation trajectory, while Fig. 3 shows four trajectories as examples. Figs. 4 and 6 instead summarize learning across all 100 trajectories, with a few individual trajectories shown only to illustrate trajectory-to-trajectory variability. We will retain this presentation in the revised manuscript rather than assigning names to the example trajectories or using them as fixed trajectories across the results.

L187: *The reduction for σ_{max} is actually quite small. Would it have been possible for uncertainty to increase?*

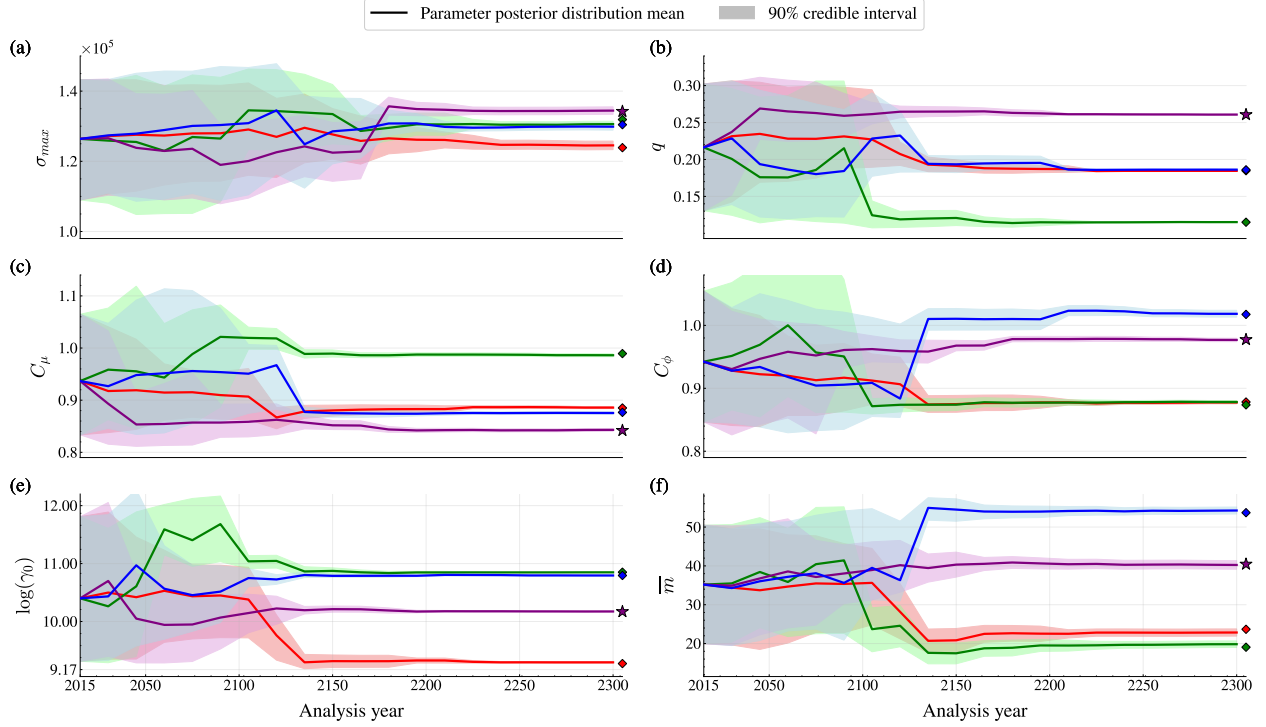


Figure 6: Future learning of MALI parameters over time for four randomly chosen synthetic future-observation trajectories. Thick lines and shaded bands denote the posterior means (*best estimates*) and associated 90% credible intervals (*very-likely* ranges), respectively. Markers on the right of each panel show the parameter values drawn from the present-day (year 2015) posterior and used to generate the corresponding trajectories. The plum-colored trajectory is marked with a star. The other three are marked with diamonds.

Yes, an increase in posterior uncertainty can occur as observations are incorporated. The across-trajectory mean credible-interval width for σ_{max} decreases throughout the revised analysis. As described in our response to Reviewer 1’s General comment 2, its across-trajectory mean 90% credible-interval width decreases from 34.2 kPa in 2015 to 2.8 kPa in 2300, an approximately 92% reduction.

L207–208: This sentence is quite unclear to me. I am not sure what you mean by “recalibrated MALI parameters”. It would be useful to define important terms in the introduction or methods section, such as “analysis”, “recalibrated parameters”, “trajectory”, and “observation”. I am still not sure whether “trajectory” is equivalent to “observation” in the manuscript.

We will define the recurring terms at first mention as described in our response to Major comment 1. For each synthetic future-observation trajectory, we first draw one parameter vector from the present-day posterior and use the 19 GP emulators of cumulative grounded mass change to obtain predictions at the corresponding analysis years. These predictions are then used to generate a time series of synthetic observations, which we call a trajectory. A trajectory is therefore not equivalent to an individual observation. “Recalibrated parameters” refers to the parameter posterior obtained after calibration using all observations available through a given analysis year.

L211–213: “Across those trajectories, SLE projection uncertainties at 2100 (Fig. 5a) remain indiscernible as information from more future years is introduced due to overlapping very-likely ranges” → This sentence is very unclear. I think the explanation is that learning is weak because there are almost no changes in ice dynamics, so each new observation provides little new information.

We will revise this sentence for clarity. The 2100 SLE projection uncertainties change little as observations are added through 2090. As discussed in Section 4.2.1, this limited learning reflects the relatively weak

forcing over much of the 21st century and the slow dynamical response of the ice sheet.

L214: Similarly, I do not think that simply having more observations is what leads to the reduction; the key is to have more distinct or informative observations. However, your assumption of independence for cumulative mass loss may hide this important point. In a Bayesian framework, what matters is new information, not simply new observations, and correlations between observations are how one accounts for whether an observation is genuinely new. If it is not, it will have a high correlation with previous observations and will therefore not change the posterior very much.

We agree that the temporal covariance affects the information provided by successive observations. However, temporal correlation alone does not determine the amount of parameter learning. This also depends on how the predicted cumulative grounded mass change responds to the parameters. As described in our response to Reviewer 1’s General comment 2, we now account for temporal correlation among the cumulative grounded mass-change observations using the Wiener covariance. In the revised analysis, the Wiener model leads to greater parameter learning than the independent-error model. We will therefore avoid describing learning as resulting simply from adding more observations.

L216: “across-trajectory mean” → Is this the mean of all observation trajectories? If so, I do not understand why it is called the best estimate. For me, in a Bayesian framework, “best estimate” usually refers to the mean of a posterior quantity.

Fig. 4 and 6: For both figures, I would show the global statistics, such as the mean and 66% CI, together with only the four trajectories shown previously, so that the figures can be followed and compared more easily. Since you do not discuss individual trajectories in detail, I do not see why all of them need to be shown. To improve the explanation of these figures, you could first explain how you compute the width of the 90% CI for each trajectory, using the four examples, and then explain how the figure generalizes this to all trajectories, which is why another interval, such as the 66% CI and mean, is obtained.

L227: “best estimate” → I suggest using “mean”, as this would make the text easier to follow and easier to link with the figure legends. The phrase “across-trajectory best estimate of widths of the very-likely range” is rather heavy and difficult to follow.

The “across-trajectory mean” is the mean of the 90% credible-interval widths across the 100 synthetic future-observation trajectories. We use “best estimate” for this mean following the terminology used in the manuscript. We will make this definition clear and simplify the wording around the credible-interval widths.

For Figs. 4 and 6, the individual trajectories are included only to illustrate trajectory-to-trajectory variability and do not correspond to the examples in Figs. 2 and 3. We will therefore retain the figure organization as described in our response to the earlier comment on tracking trajectories across figures.

L229–231: “the dominant control” → This seems a bit too strong, especially regarding ice-shelf buttressing. Grounding-line migration and changes in basal shear stress close to the grounding line are at least as important, if not more so. Moreover, Fig. 7 shows that the ice shelf remains quite stable until 2100, so I think another key mechanism could explain the importance of melting: melting plays a major role in grounding-line retreat, which reduces resistance and leads to an increase in ice discharge.

We agree that “dominant control” is too strong. The physical interpretation of the parameter-learning results will be reconsidered in light of our response to Reviewer 1’s General comment 2. We will also acknowledge that grounding-line migration and the near-grounding line stress balance may contribute.

L228–234: I do not think the explanation “As anomalies in grounded ice discharge directly contribute to grounded mass change, our synthetic observations of cumulative grounded mass change strongly inform this parameter” is fully satisfactory. C_ϕ and C_μ also affect ice discharge, since lower friction or viscosity leads to higher ice discharge, so a change in ice discharge should also inform these parameters. Physically, I would suggest the following explanation. Basal melt acts as the trigger, because it affects the magnitude of grounding-line retreat. In Fig. 7, between 2020 and 2100, the grounding line does not always migrate to the same position, which is likely due to different melt rates near the grounding line associated with different $\log(\gamma_0)$ values. This may occur relatively quickly, so the calibration can directly learn this parameter. I think this is also why learning for $\log(\gamma_0)$ is faster than for $\bar{m}$, because $\bar{m}$ controls the average melt rate and likely has less direct influence close to the grounding line. Then, for q , C_μ , and C_ϕ , some dynamical changes,

such as changes in grounding-line position, are needed to identify the appropriate parameter values. Without changes in dynamics, no new information is added relative to the 2015 observations, so little learning occurs. This may explain why these parameters respond later or more slowly than $\log(\gamma_0)$.

We agree that the connection between grounded ice discharge and cumulative grounded mass change does not uniquely explain learning about $\log(\gamma_0)$, since C_μ and C_ϕ also affect ice discharge. As mentioned in our response to Reviewer 1’s General comment 2, the revised analysis substantially changes the relative learning among these parameters, so the discussion in Section 4.1 will be revised accordingly. We will therefore not retain the original interpretation that motivated this passage.

L242–244: Yes, I think this is because $\log(\gamma_0)$ mostly affects grounding-line retreat; see the comment above.

Please see our response to the preceding comment. We will revise the physical interpretation of the parameter-learning results accordingly.

L252–253. In Fig. 7, there is also a large change in grounding-line position over this period, so I think this is also a good candidate mechanism, if not a better one.

Grounding-line migration may contribute to this behavior. We will consider it when revising the physical interpretation of the parameter-learning results.

L277–279: As mentioned above, does $\bar{m}$ not affect the average melt rate, rather than specifically the melt rate at the grounding line?

$\bar{m}$ represents the historical ice-shelf-averaged melt rate and is used to determine the bias correction δT in the melt parameterization. We will explain this. The interpretation of future learning of $\bar{m}$ will also be updated based on our response to Reviewer 1’s General comment 2.

L282: “This is might be due” $\rightarrow$ I am not sure, but is this formulation correct? Should it be “This might be due to ...” instead?

Yes. We will correct this wording if the sentence is retained in the revised discussion.

L289–290: Once the system consists of tidewater glaciers, is there no further calving? Is there only melt at the front, or are there no further changes in ice-front position?

σ_{max} governs calving of floating ice and is not applied at grounded ice fronts. As described in Jantre et al. (2024), grounded calving is not calibrated. We will clarify the treatment and evolution of the ice front after ice-shelf loss in the revised discussion.

L280–291: could the σ_{max} range be too restrictive? In Jantre et al. (2024), σ_{max} has almost no influence, but if the range were larger, we might see different frontal retreat and therefore different dynamics.

The prior range for σ_{max} is adopted from Jantre et al. (2024). It was selected from sensitivity tests exploring model configurations that produce approximately stable calving-front positions across all of the ice shelves of the Antarctic Ice Sheet. We have not evaluated a wider range applied to Amery Ice Shelf, so we cannot determine how it would affect frontal retreat or the resulting dynamics. Changing this range would also require a new MALI simulation ensemble.

L294–302: This paragraph concerns projections to 2100. Should it therefore be in Sect. 4.2.1?

We agree that the 2100-specific discussion would fit more naturally in Sec. 4.2.1. Section 4.2 will be reorganized as needed when incorporating the revised sea-level-contribution results described in our response to Reviewer 1’s General comment 2.

L297–299: “As seen in Fig. 5a, such a wide very-likely range around the posterior predictive mean of the 2100 SLE projections by 2090 would likely include 0 mm SLE, indicating that no sea-level contribution may occur by the end of the century.” $\rightarrow$ I do not see the purpose of this sentence. More generally, this paragraph feels a bit strange, as it seems to describe results that have already been shown.

The purpose of this sentence is to interpret the uncertainty relative to the magnitude of the projected 2100 sea-level contribution. In the revised analysis, the 2100 contribution remains near zero and is small relative to its approximately 3 mm 90% credible-interval width, so a zero contribution remains plausible. We will

retain this interpretation in the revised discussion, with its focus on the uncertainty.

L308–309: As mentioned earlier, this may be because new observations are being added, but they do not necessarily provide new information, which is what matters for Bayesian calibration.

Please see our response to L214. We will revise this wording so that learning is not attributed simply to the number of observations added.

L323: One point I did not fully understand when comparing Sects. 4.2.1 and 4.2.2 is the following: if some learning for 2200 projections mostly occurs around 2075, why do the 2100 projections show almost no learning?

The observations available by 2075 can constrain parameters that affect the longer-term ice-sheet response, even though the projected response by 2100 remains small. These constraints can therefore reduce uncertainty in the 2200 projection more strongly than in the 2100 projection.

L342: This is related to the previous point. Why does the 22nd century still have large uncertainties if you have shown just before that uncertainty in 2200 projections can be reduced?

These refer to different projection horizons. The uncertainty discussed here is for the 2300 sea-level contribution evaluated during the 22nd century, whereas the preceding discussion concerns uncertainty in the 2200 sea-level contribution. The two projection horizons can therefore have different learning rates. We will emphasize the different time horizons in these sections to eliminate this potential confusion.

L352–353: As above, could this not be due to grounding-line migration?

Grounding-line migration may contribute to this behavior. We will consider it in the revised physical interpretation.

L366–369: It would be useful to remind the reader which glacier area or system each of these studies focuses on.

Jantre et al. (2024) focuses on the Amery sector, while Rosier et al. (2025) considers the Amundsen Sea sector. We will add these regional identifiers in the revised manuscript.

L371: Is this not too late in terms of sea-level-rise management? This is also why I think other observations might provide an earlier signal, which would be more useful for SLR management.

The timing of learning is relevant for sea-level-rise management. As mentioned in our response to Reviewer 1’s General comment 2, the revised analysis changes the amount and timing of learning, so we will update this discussion based on those results. As discussed in our response to Major comment 4, other observables could provide earlier information about dynamical change, but we do not plan to evaluate alternative future observables in this study.

L413–416: I do not particularly like citing my own papers, but I think it would be useful to also mention Jager et al. (2024b) here. We showed that using global variables as calibration targets, such as ice discharge or cumulative ice-mass loss, provides much weaker constraints than using spatial data, such as elevation or velocity changes. This is because global targets can hide compensating errors: if a model is able to reproduce velocity and elevation changes, it should also be able to reproduce ice discharge and ice-mass loss, but the opposite is not necessarily true. For example, an overestimated ice thickness can compensate for an underestimated velocity when computing ice discharge.

We agree that a spatially integrated observable can hide compensating spatial errors. This point is already discussed in Section 4.4, where we note that cumulative grounded mass change can obscure spatial discrepancies between the model and observations. We also cite Felikson et al. (2023) for the complementary constraints provided by different spatially resolved observations. We will add the suggested reference where those points are discussed as it is indeed relevant there.

L441: “negative learning” → It would be useful to explain what this means, what can cause it, and what its consequences are.

We will explain this term. In Ceres et al. (2017), “negative learning” occurs when additional observations lead to inference in the wrong direction because the model used for inference does not adequately represent

the underlying system. We use the term here specifically in relation to model structural error, which can lead to biased or overconfident projections.

L449 : “50–100 years of observation” → Please specify “50–100 years of mass-change observations”, because this timescale might be quite different for other types of observations.

That is a good clarification. We will specify that this timescale refers to observations of cumulative grounded mass change. The stated timescale will be updated based on our response to Reviewer 1’s General comment 2.

L475–483: Here, you should discuss that, in your case, learning may occur too late for policy-relevant questions. Policymakers are likely most interested in the timing and magnitude of substantial mass loss. Performing a similar analysis with other types of observations, as mentioned above, might help address this issue. This would then raise the question of which observations are most relevant or important.

The submitted manuscript already discusses the implications of learning timescales for policy and observing-system design in Section 4.5. We will update this discussion based on the revised analysis results described in our response to Reviewer 1’s General comment 2. As noted in our response to Major comment 4, other observables could provide earlier information about dynamical change, but we do not plan to evaluate alternative future observables in this study.

Fig. A1: To me, it is unclear what each point represents.

Each point represents one SSP5 ensemble member evaluated under five-fold cross-validation. We will include this explanation in the revised Fig. A1 caption. We will also remove the predictive intervals from the predicted-versus-simulated plot because GP predictive variance (code uncertainty) is excluded from the revised analysis, as described in our response to Reviewer 1’s General comment 2.

Figs. B1 and B2 These figures do not seem to provide new information; perhaps they could be removed

We agree that these figures are redundant with Figs. 4 and 6 and will remove them from the revised manuscript.

Supporting note: a linear-Gaussian inverse modeling illustration

We use a fully analytical linear-Gaussian inverse problem to examine why, once emulator code uncertainty is removed, the temporally correlated (Wiener) error model can provide more parameter learning than the independent model (response to Reviewer 1’s General comment 2). Six parameters each drive a cumulative signal observed at the same 19 analysis years used in the study. The observation error is either independent across analysis years (the independent-levels model) or a discretized Wiener process (the correlated model), with *identical marginal variances at each analysis year*. The two models therefore differ only in their temporal correlation. The Wiener model is equivalent to using the increments between successive analysis years of the cumulative signal, while the independent model uses the cumulative *levels*, whose error variance grows with horizon. Because the problem is Gaussian and linear, the posterior credible intervals can be obtained in closed form, with no MCMC, no emulator, and no noise draw required. We consider two cases. In **Case A**, we choose six observational sensitivities that approximate, as closely as possible, the corresponding emulator sensitivities in our study (Fig. 3). Each has a sharp, localized feature, either a peak or trough, in the transient around 2180–2225, so most of its change between successive analysis years is concentrated there (Fig. 7). We refer to this as a change-like sensitivity. In **Case B**, we instead use smooth, monotone sensitivities with no localized change (Fig. 8), which we refer to as a level-like sensitivity.

The 90% credible-interval widths contract from the prior as observations accumulate (Figs. 9 and 10). Their ratio $R = W_{\text{ind}}/W_{\text{corr}}$ (Fig. 11) shows different behavior in the two cases: in Case A, the Wiener model gives the narrower interval for four of the six parameters by the final analysis year, while in Case B the independent model gives the narrower final interval for five parameters and the two models are nearly indistinguishable for the sixth (panel (e) of Fig. 10). This is consistent with the analysis presented in response to Reviewer 1’s General comment 2 in Fig. 1, where the Wiener model (scenario D) leads to greater parameter learning than the independent model (scenario C). It is also aligned with Tiedeman and Green (2013), who find that temporally correlated observation error can increase or decrease parameter learning depending on the sensitivity structure. Thus, the effect of temporal correlation depends on whether the sensitivity is more change-like or level-like.

We do not over-interpret the four-of-six Wiener advantage in Case A. We chose the six sensitivities to resemble the emulator’s, but two of them (q and C_μ , panels b and c) are only weakly change-like and correspondingly favor the independent model for parameter learning; near the boundary between change-like and level-like, the ordering is delicate. The chosen example is also deliberately idealized so that it has no emulator approximation error or MCMC sampling, and its sensitivities are only an approximation of those in the emulator. Its purpose is to explain the *mechanism*, rather than reproduce the per-parameter results of the full analysis. In the full analysis, the emulator sensitivities exhibit substantial changes between successive analysis years (Fig. 3), and the Wiener model yields greater parameter learning for all six parameters (Table 1, C vs. D).

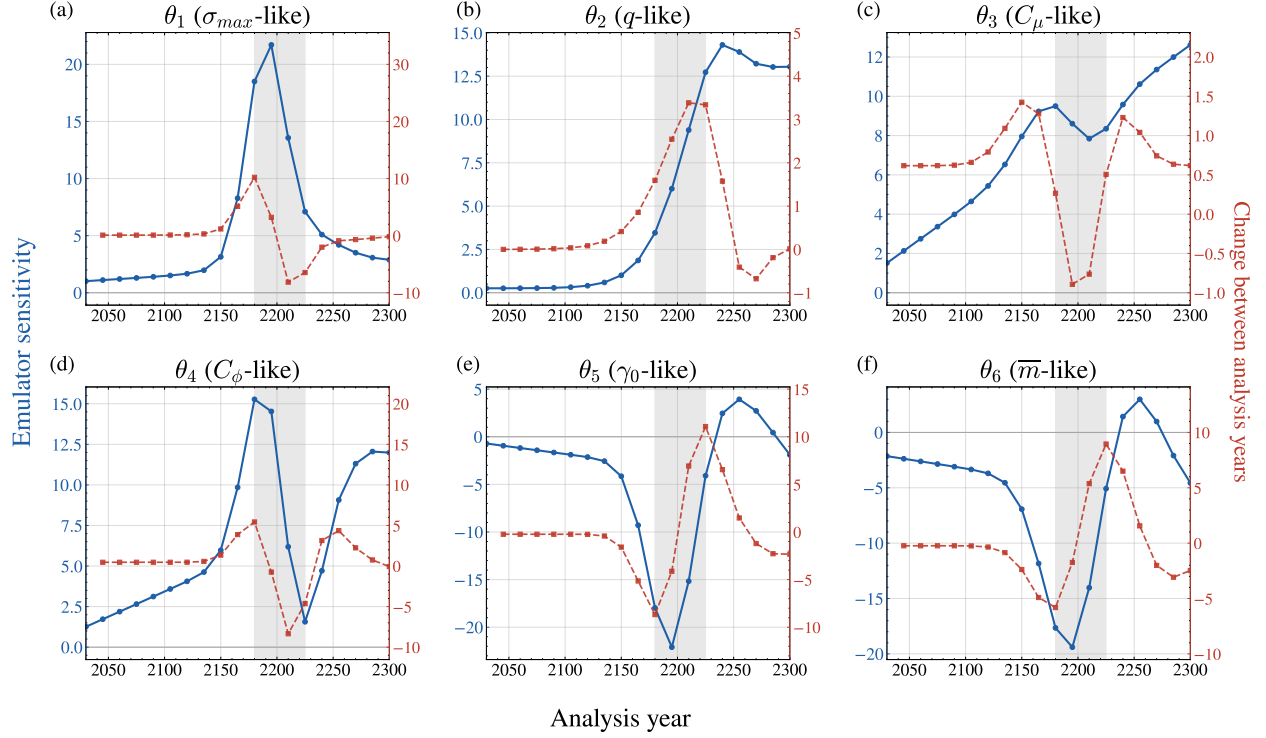


Figure 7: Linear-Gaussian inverse modeling illustration, Case A observational sensitivities (chosen to resemble the emulator sensitivities; cf. Fig. 3). For each parameter θ_j , the cumulative sensitivity (blue, left axis) and its change between successive analysis years (red dashed, right axis) are shown versus analysis year. Each sensitivity has a sharp, localized feature (peak or trough) in the transient (shaded band, ~ 2180 – 2225), so much of its change between successive analysis years is concentrated there. This represents the “change-like” sensitivity structure emphasized by the Wiener (increment) model. Panels are ordered θ_1 – θ_6 (σ_{max} -, q -, C_μ -, C_ϕ -, γ_0 -, $\overline{m}$ -like).

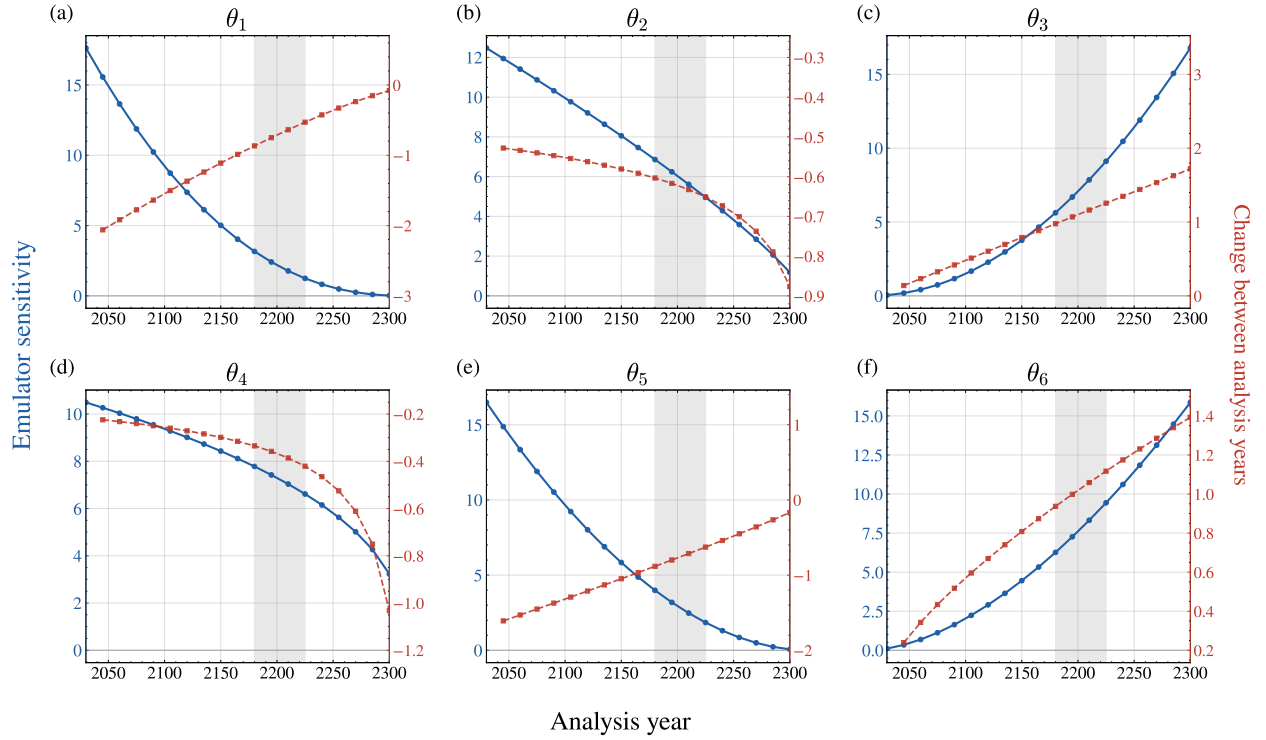


Figure 8: Linear-Gaussian inverse modeling illustration, Case B sensitivities (level-like contrast). As in Fig. 7, the cumulative sensitivity (blue) and its change between successive analysis years (red dashed) are shown for each parameter. The cumulative sensitivities are smooth and monotone, with small and slowly varying changes and no localized transient. This represents the “level-like” sensitivity structure, for which the independent model generally provides more learning than the Wiener model.

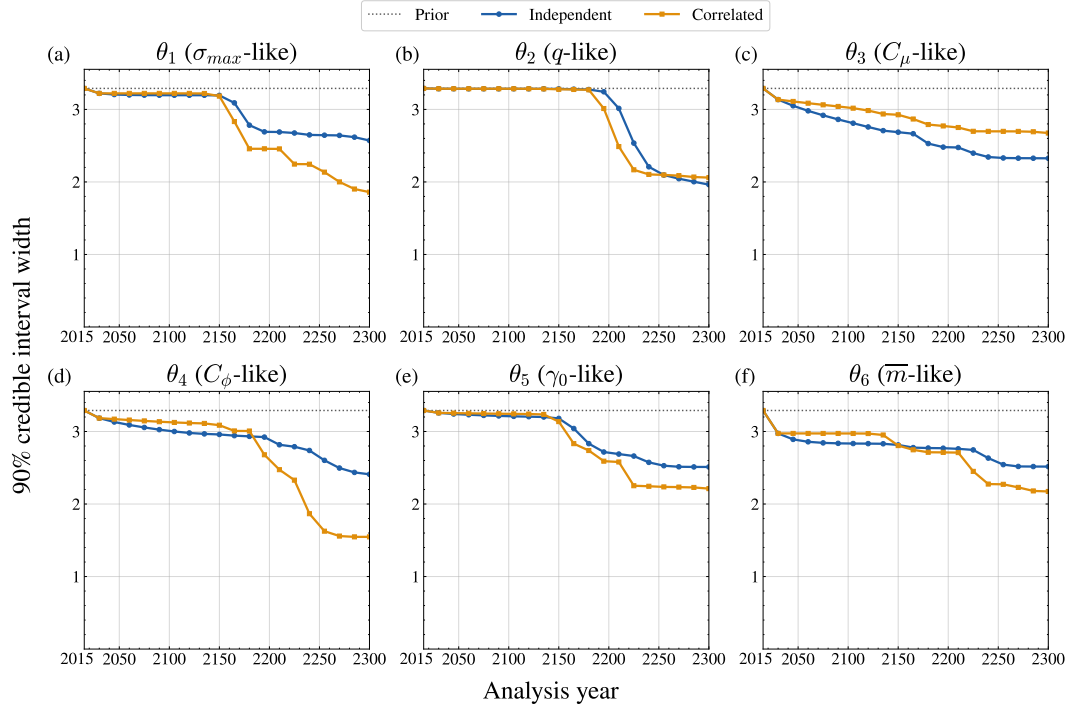


Figure 9: Linear-Gaussian inverse modeling illustration, Case A. The 90% posterior credible-interval width of each parameter θ_j is shown versus analysis year for the prior (dotted), the independent-error model (blue), and the correlated (Wiener) error model (orange). The two error models have identical marginal observation variances and differ only in temporal correlation. By the final analysis year, the Wiener model yields narrower intervals for four parameters (panels a, d, e, f), while the independent model yields narrower intervals for the other two (panels b, c).

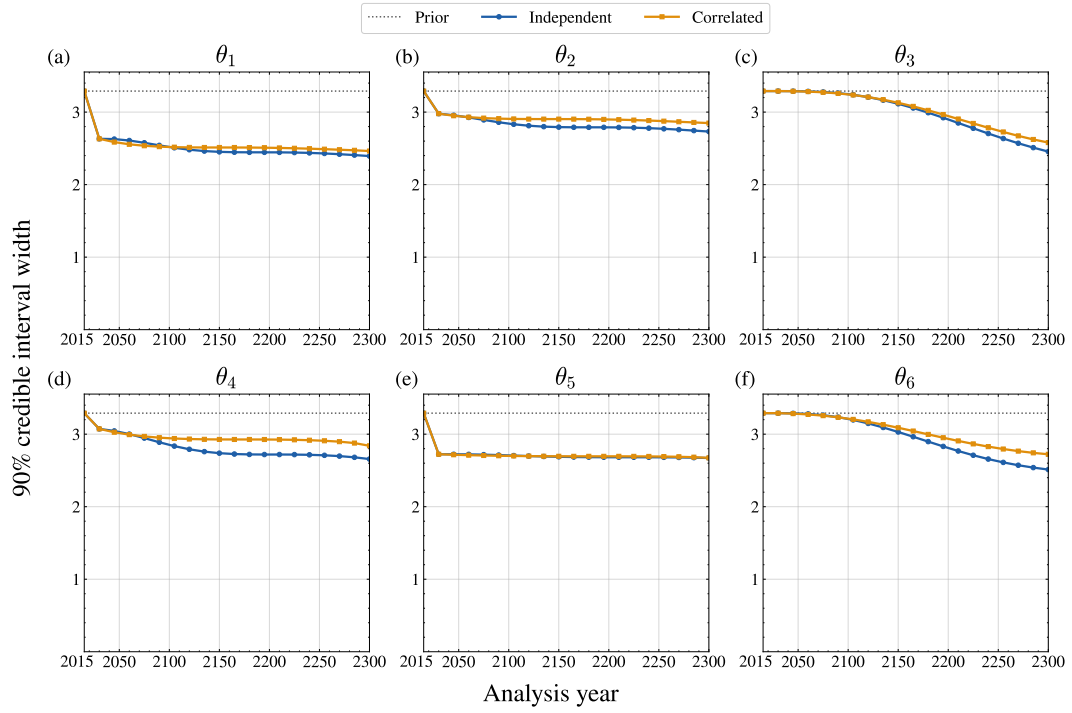


Figure 10: Linear-Gaussian inverse modeling illustration, Case B. As in Fig. 9, the 90% posterior credible-interval widths are shown for the smooth, monotone (level-like) sensitivities of Fig. 8. By the final analysis year, the independent-error model (blue) yields narrower intervals than the correlated (Wiener) error model (orange) for five parameters. In panel (e), the two interval widths are nearly indistinguishable. The Wiener model does not yield a narrower final interval for any parameter.

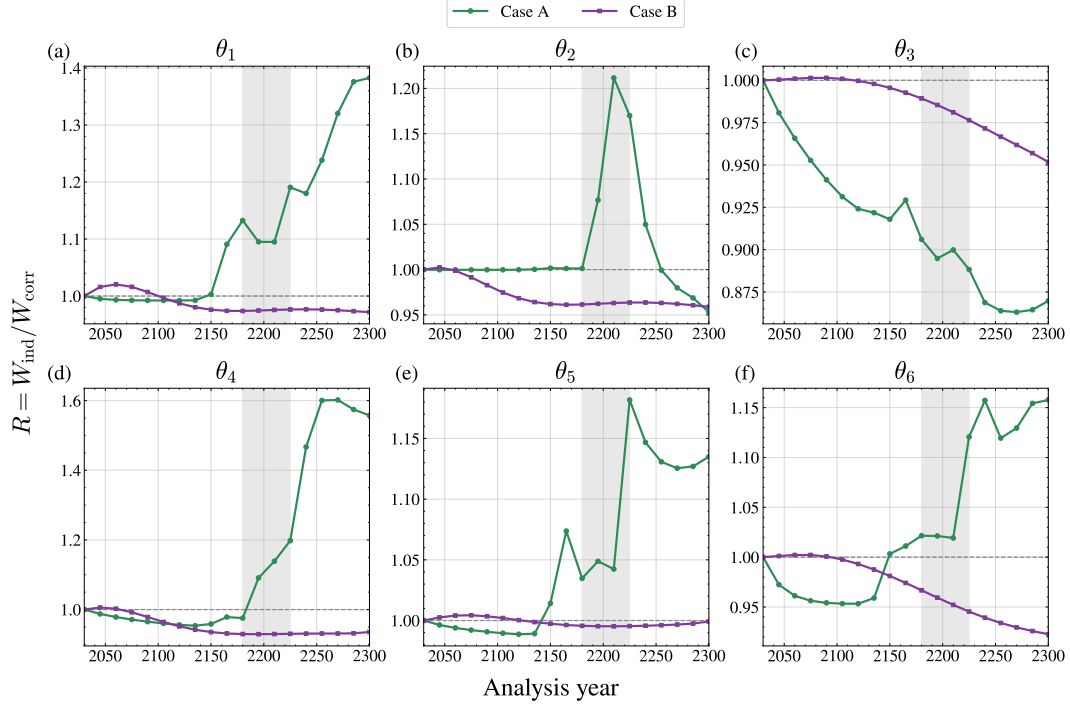


Figure 11: Linear-Gaussian inverse modeling illustration, per-parameter learning ratio $R = W_{\text{ind}}/W_{\text{corr}}$ versus analysis year. W_{ind} and W_{corr} are the 90% credible-interval widths under the independent- and correlated (Wiener) error models, respectively (Figs. 9 and 10). Values of $R > 1$ indicate a narrower interval under the Wiener model, and the dashed line at $R = 1$ indicates equal widths. For Case A (green), the Wiener model yields the narrower final interval for four parameters (panels a, d, e, f). Panel b exceeds $R = 1$ only temporarily and is below $R = 1$ by the final analysis year. For Case B (purple), the independent model yields the narrower final interval for five parameters, while the two models are nearly indistinguishable in panel e.

References

- Ceres, R. L., Forest, C. E., and Keller, K.: Understanding the Detectability of Potential Changes to the 100-Year Peak Storm Surge, *Climatic Change*, 145, 221–235, <https://doi.org/10.1007/s10584-017-2075-0>, 2017.
- Chaloner, K. and Verdinelli, I.: Bayesian experimental design: A review, *Statistical science*, 10, 273–304, <https://doi.org/10.1214/ss/1177009939>, 1995.
- Felikson, D., Nowicki, S., Nias, I., Csatho, B., Schenk, A., Croteau, M. J., and Loomis, B.: Choice of observation type affects Bayesian calibration of Greenland Ice Sheet model simulations, *The Cryosphere*, 17, 4661–4673, <https://doi.org/10.5194/tc-17-4661-2023>, 2023.
- Higham, D. J.: An algorithmic introduction to numerical simulation of stochastic differential equations, *SIAM review*, 43, 525–546, <https://doi.org/10.1137/S0036144500378302>, 2001.
- Huan, X., Jagalur, J., and Marzouk, Y.: Optimal experimental design: Formulations and computations, *Acta Numerica*, 33, 715–840, 2024.
- Jantre, S., Hoffman, M. J., Urban, N. M., Hillebrand, T., Perego, M., Price, S., and Jakeman, J. D.: Probabilistic projections of the Amery Ice Shelf catchment, Antarctica, under conditions of high ice-shelf basal melt, *The Cryosphere*, 18, 5207–5238, <https://doi.org/10.5194/tc-18-5207-2024>, 2024.
- Jourdain, N. C., Asay-Davis, X., Hattermann, T., Straneo, F., Seroussi, H., Little, C. M., and Nowicki, S.: A protocol for calculating basal melt rates in the ISMIP6 Antarctic ice sheet projections, *The Cryosphere*, 14, 3111–3134, <https://doi.org/10.5194/tc-14-3111-2020>, 2020.
- Mouginot, J., Rignot, E., Scheuchl, B., and Millan, R.: Comprehensive Annual Ice Sheet Velocity Mapping Using Landsat-8, Sentinel-1, and RADARSAT-2 Data, *Remote Sensing*, 9, 364, <https://doi.org/10.3390/rs9040364>, 2017.
- Otosaka, I. N., Shepherd, A., Ivins, E. R., Schlegel, N.-J., Amory, C., van den Broeke, M. R., Horwath, M., Joughin, I., King, M. D., Krinner, G., Nowicki, S., Payne, A. J., Rignot, E., Scambos, T., Simon, K. M., Smith, B. E., Sørensen, L. S., Velicogna, I., Whitehouse, P. L., A, G., Agosta, C., Ahlstrøm, A. P., Blazquez, A., Colgan, W., Engdahl, M. E., Fettweis, X., Forsberg, R., Gallée, H., Gardner, A., Gilbert, L., Gourmelen, N., Groh, A., Gunter, B. C., Harig, C., Helm, V., Khan, S. A., Kittel, C., Konrad, H., Langen, P. L., Lecavalier, B. S., Liang, C.-C., Loomis, B. D., McMillan, M., Melini, D., Mernild, S. H., Mottram, R., Mouginot, J., Nilsson, J., Noël, B., Pattle, M. E., Peltier, W. R., Pie, N., Roca, M., Sasgen, I., Save, H. V., Seo, K.-W., Scheuchl, B., Schrama, E. J. O., Schröder, L., Simonsen, S. B., Slater, T., Spada, G., Sutterley, T. C., Vishwakarma, B. D., van Wessem, J. M., Wiese, D., van der Wal, W., and Wouters, B.: Mass balance of the Greenland and Antarctic ice sheets from 1992 to 2020, *Earth System Science Data*, 15, 1597–1616, <https://doi.org/10.5194/essd-15-1597-2023>, 2023.
- Perego, M., Price, S., and Stadler, G.: Optimal Initial Conditions for Coupling Ice Sheet Models to Earth System Models, *Journal of Geophysical Research: Earth Surface*, 119, 1894–1917, <https://doi.org/10.1002/2014JF003181>, 2014.
- Rainforth, T., Foster, A., Ivanova, D. R., and Bickford Smith, F.: Modern Bayesian experimental design, *Statistical Science*, 39, 100–114, 2024.
- Recinos, B., Goldberg, D., Maddison, J. R., and Todd, J.: A Framework for Time-Dependent Ice Sheet Uncertainty Quantification, Applied to Three West Antarctic Ice Streams, *The Cryosphere*, 17, 4241–4266, <https://doi.org/10.5194/tc-17-4241-2023>, 2023.
- Rosier, S. H. R., Gudmundsson, G. H., Jenkins, A., and Naughten, K. A.: Calibrated sea level contribution from the Amundsen Sea sector, West Antarctica, under RCP8. 5 and Paris 2C scenarios, *The Cryosphere*, 19, 2527–2557, <https://doi.org/10.5194/tc-19-2527-2025>, 2025.

Ryan, E. G., Drovandi, C. C., McGree, J. M., and Pettitt, A. N.: A Review of Modern Computational Algorithms for Bayesian Optimal Design, *International Statistical Review*, 84, 128–154, <https://doi.org/10.1111/insr.12107>, 2016.

Tiedeman, C. R. and Green, C. T.: Effect of correlated observation error on parameters, predictions, and uncertainty, *Water Resources Research*, 49, 6339–6355, <https://doi.org/10.1002/wrcr.20499>, 2013.