Review of "Computationally efficient subglacial drainage modelling using Gaussian Process emulators: GlaDS-GP v1.0" by Hill et al.
Reviewer: Vincent Verjans

This study develops a Gaussian Process (GP) emulator to emulate output from the subglacial hydrology model GlaDS (Werder et al., 2013). In particular, the GP is trained to reproduce the sensitivity of flotation fraction output to 8 different GlaDS parameters. A principal component truncation is performed to reduce the dimensionality of the outputs to be emulated. Training is performed on an idealized glacier configuration, with a pre-specified melt input forcing. The performance of the emulator is then evaluated on 100 test combinations of GlaDS parameters, unseen during the emulator training.

This study contributes positively to efforts towards computationally efficient solutions to simulate subglacial hydrology. It also offers a promising tool to evaluate parametric uncertainty of subglacial hydrology models. This latter aspect is important, since subglacial hydrology models are heavily parameterized, with very few physical constraints on parameter values. I value positively the technical approach used for the emulator development. On the idealized configuration tested here, the emulator shows a good performance on non-training data samples. The manuscript is clearly structured and well-written. I have nonetheless a concern regarding the impact of the study. The authors have developed a subglacial hydrology emulator, but the real scientific value of this lies in the implications for areas where subglacial hydrology plays a role, many of which are provided as motivations in the introduction. As presented, both the emulator performance and the potential for uncertainty quantification are hard to interpret, because no application of the emulator is demonstrated. I detail this concern in my Major comment below, and I emphasize that this lack of impact (1) is my personal opinion, and it is the editor who decides which impact is expected from studies published in Geoscientific Model Development, and (2) does not influence my positive opinion about the quality of the work performed by the authors, but only on what more could be done. My review further includes a Minor comment regarding the quantitative evaluation of the emulator, and Technical comments aiming to improve the structure and clarity of the manuscript. Line numbers in this review correspond to the preprint manuscript.

**Major comment: Impact**
As mentioned in my introduction, emulation and uncertainty quantification of subglacial hydrology by themselves are not of great scientific interest. It is really the implications of subglacial hydrology for different fields, primarily ice flow modeling but also others listed in the introduction, that make it a critical research topic. However, none of these implications is explored here. As such, I feel like the prediction performance and the potential for uncertainty quantification from the GP emulator are not very meaningful as presented.
I note that GlaDS has been run with the Ice-sheet and Sea-level System Model (ISSM, Larour et al., 2012). As such, it should not be a big step to compute ice flow simulations (1) using the GlaDS output and (2) using the emulator output in order to evaluate the implications of emulator performance on modeled ice flow. Furthermore, it would be very interesting to see differences in modeled ice flow across the range of GlaDS parameters investigated in this study. This would really demonstrate the benefits of uncertainty quantification of subglacial hydrology when it comes to modeling ice flow velocities. Even though this study focuses on a single idealized glacier and melt forcing configuration, such ice sheet model experiments would be a great contribution to constraining ice flow uncertainty caused by subglacial hydrology.
If the authors are concerned about the length of the manuscript if such experiments are included,

I would recommend reconsidering the inclusion of the simulations of scalar quantities ($f_Q$, $T_s$, and $L_c$). In my view, these experiments are not of great relevance, as I do not see which research area would benefit from predictions and uncertainty quantification of these variables.
Again, I repeat here that the decision of sufficient impact from this study for publication in Geoscientific Model Development is ultimately a decision of the editor. I express here my personal opinion. And I emphasize that the work presented in this manuscript is of good quality, with only a single Minor comment and some Technical comments that I provide below.

**Minor comment: Quantitative evaluation**
The title of Section 6.1 is "What is the fidelity of the subglacial drainage model emulator?". In my view, this question has not been evaluated thoroughly enough. I think that simply adding a table with important evaluation metrics would be sufficient to address this concern. Evaluation metrics could be averaged spatio-temporally as well as across the 100 test simulations. It would be insightful to provide 5th, 50th, and 95th percentiles of RMSE, MAPE, coefficient of determination ($R^2$), and bias across the 100 test simulations, where these metrics are time- and spatially-averaged. In addition, it would be nice to add the same metrics but (i) for the upper and lower 30 km parts of the domain separately, and (ii) for the DJF and JJA months separately. Such evaluation metrics would give the reader a better and more quantitative appreciation of the performance of the GP emulator. Finally, for each of these metrics, I recommend also providing between parentheses the same metric but computed on the training data. This would be insightful to evaluate the potential degradation of the GP emulator performance when used on inputs unseen during training.

**Technical comments**

- General (1): The authors make an excessive use of parentheses throughout the text. In my view, parentheses should only be used to provide additional non-essential details in the text. I recommend that the authors clean up their parentheses by making more sentence separations instead of overloading single sentences.

- General (2): Throughout the manuscript, the authors use both the terms "inputs" and "parameters" to refer to the same notion: the parameters of GlaDS passed to the emulators. To avoid any confusion, a single term should be used consistently in the entire manuscript.

L5
Replace "construct robust" by evaluate uncertainty in.
L6
uncertainty quantification.
L14
"of the water pressure variance": it is unclear if this refers to spatial variance, temporal variance, and/or variance across the samples of the parameter space.
L15
I believe that the mention to observational data is misused here, as no observational data is integrated in this study.
L25
"well-established": I understand what the authors mean here. However, this wording is misleading, because although there is consensus about the existence of an influence of subglacial hydrology on ice flow, this influence remains highly uncertain.
L26
Replace flow by sliding.

L31-36
This sentence is too long, and I do not know what "which" (L36) refers to.
L36
Replace large by high-dimensional.
L52
Typo: "GP emulators we develop".
L53
I suggest this definition for flotation fraction: ratio of water pressure to ice-overburden pressure.
L55
Please explain here the meaning of "global sensitivity indices".
L63
If possible, please use another word than "emerging".
L69
Please provide units of variables.
Eq. (1)
Although obvious, please define $g$.
L77
" each of these quantities defines a two-dimensional, time-varying field": this is confusing because I believe that both $z_b$ and $p_i$ are not time-varying.
L83
Please specify: For details about GPs
L85
I do not understand what is meant by: "in terms of the proportion of output variance corresponding to each GlaDS parameter". Please clarify.
L91
Rephrase: Let $\boldsymbol{y}_i$ denote the vectorized model output of all variables (...).
L94
Typo: "which are not a part of".
L104
Add a comma after $\boldsymbol{\theta}$.
L110
Refer to $\mu(\boldsymbol{x})$ after "mean function".
L110
"to set the mean to zero" should be to set the prior mean to zero. In the following sentences, it is also important to emphasize that it is only the prior mean that is zero.
Eq. (5)
This should be $y_p|\mathbf{Y}, \theta, x_p$
L124
I recommend being more specific here: (...) contains the pair-wise covariances between $x_p$ and each entry of $x$ (...).
L126
Specify: The prediction mean
L127
Refer to Eq. (4) after "covariance function".
L136 and L137
Replace "will" by would.
L138
Typo: "a variety solutions".

L148
Please specify here that Eq. (8) assumes uncorrelated errors. This may not be entirely valid in this case.
L152
"can be viewed as": this wording is inappropriate, because it is a dimension reduction by definition.
L162
The authors can also invoke the orthogonality property of the PC decomposition to motivate their univariate approach.
L169
Replace "permissive" by flexible.
L169
"variations in the principal components that tend to be smooth with respect to the input parameters": why is that? I would expect a strong sensitivity of GlaDS to some of its parameters, even in the PC subspace. Could the authors please clarify this statement?
L173
How many GP realizations are sampled?
L179-181
I am not sure to agree here. As I understand it, each univariate GP is fitted to a single series of PC coefficient, regardless of the number $p$ of PCs retained. The number of parameters scales linearly with $p$, being $p(d+1)+1$. However, the amount of data used for fitting also scales linearly with $p$, because increasing $p$ by 1 implies that one more series of PC coefficients is used. As such, I do not see why "a simpler model with fewer PCs and therefore fewer hyperparameters to estimate is desirable as it will have less prediction variance (i.e., less tendency to overfit)". On the other hand, I believe that using an increasingly high number of PCs would imply increasingly many GPs fitted to low-variance component of the GlaDS output, which can be regarded as noisy features of GlaDS results rather than dominant components of the variability.
Section 3.1
In general, I think that more details are needed in this Section.
L192
Add one sentence to explain what the K-transect is.
L192-193
"adjusted from 0 m" and "increased to 40–1560 m": are these adjectives with respect to the SHMIP configuration? If so, please specify.
L196
Please put the basal melt rate imposed into a glaciological context. For example, how does it compare with basal melt rate estimates in Greenland, or with the SHMIP forcing?
L199
" following a moulin density that varies with elevation computed from a satellite-derived supraglacial drainage map": is it possible to provide the formulation of the moulin density as a function of elevation?
L200
"within each sub-catchment": this is not explained.
Figure 1
Is it possible to add the melt rate using the right y-axis in Fig. 1a? What does the color scheme represent in Fig. 1b? Is it possible to indicate the moulin locations in Fig. 1d?
L203
Replace "posed" by configured.
L204

The variable $x$ is already used to denote the input to the GPs. Please do not use the same symbol for two different variables.

L216

"Following the vocabulary of Higdon et al. (2008) and Verjans and Robel (2024)": this is not needed here.

L222

Concerning the parameter ranges, the ranges provided in Table 2 are most likely not intuitive to a majority of readers. I recommend adding a column in Table 2 specifying the ranges of parameter values used in previous studies focused on uncertainty quantification from GlaDS parametric uncertainty (e.g., Brinkerhoff et al., 2021).

L223

"flotation fraction $f_w < -10$": in principle, any $f_w < 0$ is nonphysical because it implies $p_w < 0$. Are all these simulations rejected from the training data? And/or is the GP constrained to predict $p_w > 0$?

L234

As I understand, the test data do not include any extrapolation beyond the parameter space used for training. Therefore, it should be mentioned here and in the Discussion that the extrapolation capability of the GP emulator has not been evaluated.

L242

Please change this sentence to "In addition to emulating the spatiotemporal flotation fraction, (...)".

L277

Remove "small".

L279

"perhaps since the input space has been explored more thoroughly": I do not think this is the case. In my view, more PCs are needed simply because the rank of the output space increases. For example, if a single simulation is run, it is fully characterized by a single PC. As more simulations are included, the number of PCs required to fully characterize the outputs increases, and thus the number of PCs to characterize a given % of the output variance also increases.

L284

"only the absolute value, not the sign, of the PC basis vectors should be interpreted": I disagree. Opposite signs indicate opposite phasing of variability. It would be more correct to say that the sign of any given PC basis vector is arbitrary, but only looking at the absolute value would be wrong.

Figure 3

Specify if the lines show the mean or median of RMSE and MAPE taken across the test simulations.

Figure 4c,f

I think that the presentation of the 95% prediction intervals is both unclear and misleading. Firstly, I understand that the RMSE and MAPE boxplots show the errors averaged in both time and space. But for the 95% prediction intervals, do the boxplots show the entire population of 95% prediction intervals taken at each grid cell and each time step of each test simulation? Secondly, this Figure suggests that broad prediction intervals are a bad thing. However, the purpose of a prediction interval is to communicate about the uncertainty in the output. Thus, it is a good thing that prediction intervals are broader for cases with high RMSE (i.e., the simulations with low PC numbers in this Figure). This means that the true GlaDS value may still lie within the 95% prediction interval despite the larger error in the mean estimate. For this reason, I recommend to show the percentage of GlaDS values falling outside of the 95% prediction intervals in Figure 4c,f, rather than the 95% prediction intervals themselves. If the GPs are well-calibrated, this percentage should be 5%.

Figure 4d,e,f
Please mention in the caption that the x-axis uses a logarithmic scale.
Figure 4 caption
"Black circles indicate the total integrated prediction uncertainty": I do not understand this.
L320
I find that it is worth mentioning that the RMSE for the three GPs is very similar for the late-September melt event, and I suggest to provide a succinct explanation of why this is.
L321
Typo: "reduces by the height".
L335
Table C1 should be referenced here.
L337
Please explain why $f_w > 2$ is considered unrealistic.
L343
"suggesting the emulator has reasonably accounted for basis truncation error": please note that this also suggests that the GP can correctly estimate uncertainty due to interpolation towards unseen parameter values.
L345
Add comma after "spring".
L345
"the mean prediction significantly overestimates flotation fraction": this suggests that the GP tends to further amplify the unrealistic GlaDS output. Please mention this explicitly.
Table 3
Specify: Single GlaDS simulation. L353 and Figure 8
Please use coefficient of determination ($R^2$) as an evaluation metric, rather than the squared correlation coefficient ($r^2$).
L350-359 and Figure 8
These comparisons are misleading, because the GP emulator has been trained to reproduce $f_w$. It is impossible to know what the performance of the GP would be if it had been trained to reproduce $\phi$ or $N$. The discussion here should be rephrased as an evaluation of the error introduced by the conversion from $f_w$ to $\phi$ and/or $N$, rather than "indicators of GP prediction performance" or "different prediction skill" (L355).
Figure 9 column c
Same comment concerning the prediction intervals as for Figure 4c,f.
L368
"Based on RMSE, MAPE and bias": I am not sure that these performances can be compared so easily from Table 4. For example, RMSE and bias have different units for the three variables. MAPE could serve as a better comparison basis, but $T_s$ has been log-transformed, and MAPE may not be representative of the error on $T_s$ itself. In addition, as mentioned by the authors, the MAPE values depend on the degree of variability in each quantity and on the values themselves (e.g., the percentage error when $\log T_s$ is 0 tends to infinity). Finally, when considering the ranges of values provided in Table 4, it is not clear to me that "the channel discharge fraction emulator has the best performance".
L389
"Sensitivity indices for the flotation field are defined as a variance-weighted sum of the sensitivity indices for each principal component": This one-sentence explanation is not clear to me. If possible, I recommend providing the mathematical formulation for the sensitivity indices. That is, which formula is used to compute the values shown in Figure 10? Furthermore, the difference between

first-order and total sensitivity indices should be explained.
Figures 10 and 11
Some whiskers extend beyond the value of 1.0. While the value of 0 can be intuitively interpreted as no sensitivity, what do values >1.0 mean?
L418
"prediction RMSE is $< 20\%$ of the ensemble standard deviation": across the 100 test simulations?
L422
"PC truncation RMSE on the test set for the reference model with 8 PCs is 0.034, while GP prediction RMSE is 0.054, suggesting the PC truncation error contributes more than half of the prediction error.": I think that this statement requires a more thorough justification, and I am also not sure that I agree with the authors about it. First, does the 0.034 value correspond to the case of 8 PCs for the curve 256 simulations in Figure 2a1? If so, please refer to Fig. 2a1 in the text. Second, Figure 4 shows that there seems to be a baseline RMSE of the GP predictions of about 0.05, which does not decrease when going from 7 to 11 PCs. On the other hand, the PC truncation error must decrease when going from 7 to 11 PCs. As such, this indicates that there is a balance between (1) using only the first few PCs that seem to be relatively easy to predict for the GP, and (2) including low-variance PCs that allow to reduce the truncation error, but that seem to be harder to predict for the GP. As a consequence of this balance, the baseline RMSE stagnates at 0.05. But saying that "Of the two error sources, PC truncation error is the larger contributor" is misleading. I believe that if more PCs had been included, PC truncation error would decrease, but GP error would increase. Thus, this conclusion seems to be due to the choice of truncating at 8 PCs, rather than an inherent attribute of the GP. At least, this is how I understand the results. I would welcome any thoughts from the authors about this.
L434
"$5 - 10 \times 10^4$ time steps": why such a range? I thought that all simulations had been performed over the same time period and with the same temporal discretization.
L446
"The impact of large errors in predicting the spring pressure maximum is also reduced for ice-flow modelling applications": I disagree with this statement. For example, Fig. 4d (magenta curve) and Fig. S4c of Verjans and Robel (2024) show that the highest ice flow velocity errors due to the subglacial hydrology emulation occurs in the spring pressure maximum (their Fig. 4d) and at the ice velocity peaks (their Fig. S4c), which generally coincide with the spring pressure maximum. So, it is impossible to verify this claim from the authors if they do not actually compare ice flow model realizations forced with the GlaDS versus the GP output.
L454-455
" the model of Verjans and Robel (2024), who report squared correlations ($r^2$)": this is not correct. Verjans and Robel (2024) report the coefficient of determination ($R^2$), which is not the same metric as $r^2$.
L474
"since PC truncation typically preferentially dampens high-frequency variations": I disagree with the authors here. PC truncation selects the components of variability with maximum variance. If most of the variance lies in high-frequency bands of the spectrum, there would not be any damping of high-frequency variations. Whether PC truncations dampens high-frequency variability or not depends on the power spectrum of the data.
L477-479
It is also important for resolving sub-annual ice flow variability.
L487
I believe that it would be relevant to add a sentence here about the propensity of GlaDS to produce

nonphysical output.
L493
Citing Verjans and Robel (2024) here is misleading, because their emulator is transferable to different domains or melt inputs, i.e., the opposite of the sentence given here.
L521-528
I found this entire paagraph a little vague and hand-waving. I recommend that the authors focus on the current work and future developments. For example, why discussing emulator predictions of global mean sea-level rise? This is clearly not the focus of this study.
L542
Typo: "is is".
L548-549
"fully Bayesian time-dependent calibration to provide observationally constrained distributions of subglacial drainage variables": I do not understand what the authors mean here.
L561-562
Add regimes after "laminar and turbulent".
Eq. A3
To make the notation less clumsy, I suggest replacing the third and fourth terms on the left-hand-side by $\frac{\partial h_s}{\partial t}$.
L603
Please specify: of a multivariate Normal distribution.
Eq. B3
For this equation to be valid, there needs to be an additional constant term on the right-hand-side.
L616
Typo: "includes" should be include.
L625
Typo: "by condition" should be by conditioning.
Eq. B5
$\theta$ should be boldfaced.
L629
then sampling from equation (B5)
L634
a major benefit of using
L635
Can the authors please remind here what are $p$ and $d$ so that the reader does not need to go back tp the main text?
Figure C2
It seems to me that some of the Markov Chains are not well-mixed, although it is hard to tell from the scale of the y-axes. Did the authors compute convergence diagnostics? I recommend providing R-hat values and effective sample sizes (Gelman et al., 2013).
Figure C3
These figures should be shown in two dimensions rather than 3 for better clarity.

## References

Douglas Brinkerhoff, Andy Aschwanden, and Mark Fahnestock. Constraining subglacial processes from surface velocity observations using surrogate-based bayesian inference. *Journal of Glaciol-*

*ogy*, 67(263):385–403, 2021.

Andrew Gelman, John B Carlin, Hal S Stern, David Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Eric Larour, Helene Seroussi, Mathieu Morlighem, and Eric Rignot. Continental scale, high order, high spatial resolution, ice sheet modeling using the ice sheet system model (issm). *Journal of Geophysical Research: Earth Surface*, 117(F1), 2012.

Vincent Verjans and Alexander Robel. Accelerating subglacial hydrology for ice sheet models with deep learning methods. *Geophysical Research Letters*, 51(2):e2023GL105281, 2024.

Mauro A Werder, Ian J Hewitt, Christian G Schoof, and Gwenn E Flowers. Modeling channelized and distributed subglacial drainage in two dimensions. *Journal of Geophysical Research: Earth Surface*, 118(4):2140–2158, 2013.