tc-2017-268 response to referees (with revisions), 30th April 2018.

Again, we would like to thank the referees and the editor for their feedback. We have now made our proposed revisions which we feel has resulted in a much improved manuscript. Below, we detail (in red) all of the revisions we have made to the manuscript in response to the referee comments. For complete transparency, we have kept all of the text from our final response document submitted on 6th April 2018 (FinalResponse_tc-2017-268.pdf) which includes the referee comments in black and our responses in blue. As well as our submission of the revised manuscript, we have also attached a mark-up document at the end of this document which details all of the revisions made. We look forward to the response from the editor at your earliest convenience.

We thank all of the referees and the editor for taking their time to read our manuscript and provide us with some very useful feedback. We received three reviews, two of which were overwhelmingly positive and one of which was largely critical of the manuscript; we greatly appreciate all of the comments. All referees clearly share our appreciation of the importance of this area of research, and as such, we hope our detailed responses below will enable us to contribute to this subject. We have taken care to address all comments, especially to those that require clarification or are critical for which we provide more detailed responses with clear justifications. We list below all of the referee comments in black followed by responses in blue. Any proposed revisions are highlighted with **bold** text. We look forward to the response from the editor at your earliest convenience.

Anonymous Referee #1

This paper presents a truly comprehensive investigation into the performance of a set of glacio-hydrological models of varying complexity. Such models basically take climate and topographic data and calculate river discharge via model calculations of distributed glacier melt, linked to overland and river flow models. The models investigated here are all at what might judged as the simpler end of model configurations used within glaciology, as the melt models are variations on the well-established temperature index approach, and the runoff models are all reservoir models, again of varying complexity. The key advance in the paper is the use of a comprehensive range of output measures used to judge model performance, and within that, the adoption of a sophisticated estimate of the limits of acceptability for each performance test based on uncertainties and errors within these datasets due to intrinsic errors or errors induced by sampling, spatial scale and the like. This is a powerful advance over most previous studies, which have used a limited set of performance measures, normally judged with rather simple statistics for model agreement, such as correlation or the various mean error statistics available. I think the argument made that such simplistic model/data comparison approaches have limited glaciological studies in comparison with other model disciplines, such as hydrology, is powerful, although it has been addressed occasionally within glaciology before. Rye et al. 2012, for instance, show how there is a need for multi-objective optimisation within melt modelling, as no single measure of model fit is adequate to fully capture the performance of any given model, and different outcomes, in terms of longer-term mass balance predictions are possible with equally acceptable models. The methodology used by Rye et al., and the overall scope of their investigation is different to the current study, but the over-arching aims seem very similar.

The paper sets out a clear methodology to develop the limits of acceptability for each 'test' dataset. As examples, the LOA as defined for ice melt include the need to allow that model-based estimates of melt when compared with in-situ stake ablation measurements can only be expected to be as close as the actual spread in melt over the equivalent node area within the model. This spread was calculated using high resolution terrestrial LIDAR scans made during the stake measurement

campaign, and used to define 95% confidence bounds around the stake measurements. MODIS-based (and therefore 500m spatial resolution) snow cover estimates were transformed to use within the 50m resolution DEM of the study site with a monte-carlo approach in which for each MODIS snow pixel the relevant number of 50m snow pixels were distributed randomly 1000 times to generate confidence bounds around the snow cover-elevation measure used as a model test. Such processes were adopted for 33 data 'signatures'.

The comparison exercise was based on the use of three variations in temperature index model (of differing complexity), run with each of three runoff models, giving 9 model configurations. Each was then run with a randomised range of parameter values, yielding 45,000 calibration runs to be judged against each of the 33 'signatures'. This set of results is evaluated both in terms of measuring the various models' performance against the range of signatures, and also in terms of the signatures' discriminatory power. Ultimately, no single model nor set of parameters met all of the LOA across all the signatures, but the approach was capable of showing the 'trade-offs' within model configuration/parameterisation in terms of the different patterns of acceptability identified, and it also showed that additional complexity did not always lead to improved acceptability.

This might be seen as a limitation of the current study. It may be that such a comprehensive analysis of a range of models and data signatures would lead to a different 'specific' conclusion in terms of the best-fit model, or the most powerful discriminatory data, when applied to a different glacier system. I do feel, however, that whilst such a complex study as this one is not going to be possible for every model application, a more sophisticated methodology to evaluate model fit is important within glaciology. I think a key specific outcome of this study (which to my mind merits publication by itself) are the descriptions and methodologies use to generate the limits of acceptability for the range of data available. These will form a useful resource for future studies which may adopt this, or similar, methodologies. Additionally, as well as potentially improving the discriminatory power of models and our ability to discriminate between different model performance, work such as the current study, and a few earlier papers, will also allow a better understanding of the uncertainty inherent in model predictions. I think this is a careful and extremely thorough study, and I strongly recommend it for publication. I hope it gains impact and this type of more sophisticated evaluation of model performance gains traction within the glaciological community.

The paper is also commendably well written. I have very few specific corrections which I feel should be made.

We greatly appreciate the positive and encouraging comments from the referee and echo the hope that this manuscript will stimulate further exploration of model evaluation approaches within the glaciology community and that it will provide a useful resource for those wishing to do so.

We agree that study of Rye et al. is somewhat different in overall aim to this study, but both were undoubtedly motivated by similar curiosities about model selection, model evaluation metrics and prediction uncertainty which warrants its inclusion in the introduction. **Accordingly, in our revised manuscript we will add additional text at the end of P4 L4 describing their findings.**

We have now added the following additional text (see P4 L10 in mark-up document) to the introduction which details the study of Rye et al. (2012) as an example of a multi-criterion approach to identify structural inadequacies in a distributed surface mass balance model:

"Rye et al. (2012) applied such an approach to 5 optimise a distributed surface mass balance model of two glaciers in Svalbard. They used ablation stake data to define three different features of the observations including mass balance at the stake locations, long term mass balance trend and mass

balance gradient. Using a multi-objective optimisation procedure, they identified structural inadequacies relating to how the mass balance gradient was simulated."

P2 L10 '. . .adhere to, . . .' (insert comma)

P3 L22 ' therefore we' (delete comma)

P3 L24 'definition imperfect' (delete comma)

P5 L15 change 'on' to 'of'

P9 L 25 delete 'an' before 'can'

P15 L1 Suggest rewording to 'As in may glaciated catchments, topography controls spatial temperature gradients to a large extent'.

P15 L7 Suggest change 'shallow' to 'reduce'

P18 L 6 'curves provide' (delete comma)

P24 Figure 8 caption or key needs to include explanation of the dots on Fig 8a P40

L16 'all but two of' (delete both commas)

Thank you for bringing these issues to our attention. **All of the above edits will be addressed in the revised manuscript.**

All of these edits have been undertaken as suggested.

Reference Rye, C.J., Willis, I.C., Arnold, N.S. and Kohler, J., 2012. On the need for automated multiobjective optimization and uncertainty estimation of glacier mass balance models. Journal of Geophysical Research: Earth Surface, v. 117, doi:10.1029/2011JF002184

David Loibl (Referee #2)

The study presents an innovative approach to evaluate the performance of glaciohydrological models. Basing on data from Virkisá river catchment in Iceland, the study demonstrates a framework to constrain the acceptability of results from different models against observational data. Different setups combining relatively simple glacier melt and discharge models are tested for their capability of reproducing measurement data results of melt, snow cover, and river runoff. Such a comprehensive framework for model evaluation is certainly an important contribution. I got the impression that the setup presented manuscript was designed very thoughtfully. The manuscript is also very well written, with well-done figures illustrating many key aspects.

Nevertheless, I see three major weaknesses of the manuscript in its current form which need to be addressed before publication: (i) The weak precipitation data set, (ii) patchy result data, and (ii) that no scripts or technical details are presented.

We appreciate these overall very positive comments from the referee, particularly with regards to the quality of the writing and presentation, as also noted by Anonymous Referee #1 and the Editor, of which we did indeed spend considerable effort in presenting in as clear and transparent way as possible. We also appreciate the clarity with which the referee has detailed their suggested weaknesses for improvement of the manuscript. We will respond to these comments sequentially as laid out by the referee.

(i) The AWS data used to force the models does not contain information on snowfall, and even rainfall with consecutive three above-freezing days was used. The manuscript provides few information of how much data is actually lost though this procedure (s. also ii), but I guess it is quite a lot in Iceland. These measures will certainly have substantial effects on glaciological-hydrological outputs and may account to some extent for the models inability to calculate winter melt. However, the authors use the data very thoughtful and have made quite some effort to ensure validity. Ideally, the framework should additionally be tested in a setting with high-quality AWS/snowfall data to constrain specific effects originating from input data characteristics. Arguably, this is beyond the scope of this manuscript; Nevertheless, I find it very important to discuss in more detail to what extent the strengths and weaknesses identified for individual model setups might also be affected by shortcomings in input/observation data.

We completely agree that the input data has a major control on model simulations and it is therefore important to stress potential weaknesses that could arise in the simulations as a result of deficiencies in the driving data. Although we did hint at these deficiencies in our discussion (P38 L5), particularly in relation to the ice melt and snow distribution signatures, we agree that we should have provided a more comprehensive discussion of this, particularly in relation to the river discharge simulations which are undoubtedly also susceptible to errors induced by deficiencies in precipitation data. **Accordingly, for the revised manuscript we will provide additional discussion of this aspect.**

In conjunction, we intend to provide additional analyses of the bias-correction procedure used for precipitation (in response to comments from Anonymous Referee #3), which will include details of lost data due to freezing days. Accordingly, the additional discussion points will draw on the results from these analyses so that weaknesses in model simulations can be better related to weaknesses in the driving data.

We have now added considerable text to the discussion to emphasise the importance of the driving climate data, particularly in relation to precipitation for which we don't have any snowfall data (see P39 L17 onwards in mark-up document). In particular, we emphasise that:

- "there were fewer observations during winter months and none at all before 2009"
- "while the bias-corrected precipitation time-series was well correlated over a three-day time-step, it was not at an hourly time-step"
- "precipitation observations were all collected at the bottom of the catchment and therefore driving precipitation data at the top of the catchment are less certain"

In the previously submitted manuscript, we already discussed how these errors could explain some of the simulated biases for the ice and snow signatures. We have now added to this with an additional example of how these errors could further propagate to the river discharge signatures:

"Furthermore, given the strong coupling between snow, ice and river runoff, deficiencies in capturing the snow and ice signatures could also propagate through the hydrological representation of the catchment. For example, one could imagine how errors in the spatial distribution of snow could perturb the timing of runoff through the catchment given 20 that snow distribution influences the behaviour of the semi-distributed runoff-routing routine employed in the GHM. Such perturbations are likely to impact the ability of the GHM to capture the full range of river discharge signatures."

We finish, by reiterating the importance of the driving climate data:

"Accordingly, it is important to stress the influence that biases in the driving climate data could have on the model acceptability across the different signatures."

We also noted that in section 2.4.1 of the original manuscript, we speculated that, "poor replication of the timing of hourly rainfall events should not influence the GHM's ability to capture the river discharge signatures". While we still feel this is true for shorter-timescale signatures given that we have complete rainfall data for 2013 and 2014, the above statement about the propagation of error from the ice and snow representation in the GHM through to the river discharge simulations somewhat contradicts this. As such, we've reworded this statement to read: "…poor replication of the timing of hourly rainfall events should have minimal influence on the GHM's ability to capture the river discharge signatures" (P14 L5 in mark-up document).

(ii) Result data tends to be patchy, making it hard for the reader to get the broader picture. For instance, Fig. 13 only shows modelling results for May, Fig. 14 only the year 2013, Fig. 16 only selected months, etc. I totally understand why these selections were made. However, I strongly recommend to add full input and output datasets as supplements to show that the examples are no cherry picking, and to allow readers to see the greater picture.

Yes we understand your concern here. To be clear, we made these selections rather than displaying the full time-series for two main reasons: i) it helps to focus the reader on those aspects of the hydrograph being analysed in the text; and ii) differences in model behaviour are more difficult to visualise when the simulation data are compressed in the graph due to displaying the full time-series. Please also note that these selections were made because they provide ideal periods to analyse different aspects of model behaviour. For example, May 2013 was chosen for Figure 13 because it is a period with almost no rainfall at the start of the melt season, and so differences in the river discharge simulations can be considered a function of the melt behaviour of the model rather than the rainfall-runoff behaviour. We could have chosen 2014 (see similar enhanced melt at start of melt season in the figure included below), but because of the presence of a number of rainfall events, we used 2013. In contrast, Figure 16 include periods where melt is insignificant compared to rainfall, and which include a range of peak river discharge magnitudes. Nevertheless, we completely agree that, for complete transparency, we should include the full input/output time-series for these simulations. **Accordingly, as part of the revised manuscript we will include an additional section in the appendix with these figures.**

We have now included these additional figures in the Appendix (see P51&52 of the mark-up document). These include the watershed total precipitation, average temperature and incident solar radiation data used to drive the models over the complete observed river discharge observation period. We have also included the complete simulations of snow melt, ice melt and river discharge (plotted with the observations) using the different GHM configurations.

Additionally, we make the reader aware of these figures immediately after referencing Figure 13 in the main body of the text (P30 L7 in the mark-up document).

(iii) Intuitively, I'd say the term 'framework' suggests that the aim is broad application, and I think the conclusions chapter encourages application and further testing of the LOA framework. However, the underlying structures and algorithms remain vague. If the aim is to let other scientists follow this approach, I suggest providing more detail on how to set up a LOA framework. As an open science and open source enthusiast, I personally would like an open git repository where interested persons could access the code and maybe help developing it further; This would certainly boost the

resonance to this good work. In combination, (ii) and (iii) render it impossible to reproduce your analysis in the current state of the manuscript.

We certainly share your enthusiasm for being open with data, publications and computer code, and given that this work is funded by a Natural Environment Research Council (NERC) studentship, all aspects of this work come under the NERC data policy (see here: http://www.nerc.ac.uk/research/sites/data/policy/) which requires all data to become open within 2 years and code to be preserved for others to use. We agree that git would be an ideal repository, however we are mindful that the code written for this study is completely tailored for our study basin and its data format. To produce a usable set of open-source scripts of the code used to calculate the LOA around the river discharge signatures (for example) would take some considerable effort. That is not to say that this is something we won't do in the future, just that if/when we do this, we'd like to do it properly and arguably, this would warrant a publication in itself. We are also surprised that you feel the underlying structure and algorithms remain vague. We feel we have been as open as we can be about the methods used. For example:

- GHM: All model equations are well established and either referenced or written in full (P9 L11 onwards).
- Driving data: all data sources have been referenced where possible, and any bias-correction procedures have been referenced and/or detailed (sections 2.2 and 2.4).
- Signatures and limits of acceptability: All signatures are listed in Table 1 and referenced in the text (e.g. P19 L18) All derived LOA are also shown in Table 1 and methods to derive these are either referenced or explained in the text (section 2.5).

Of course, as with any publication, we would also be happy to receive correspondence from other researchers who may have further questions about the methods used for which we would provide advice and share code where possible**. We feel it is important to make this clear to the reader and as such, in our revised manuscript we will include some text at the end with regards to sharing of the code.**

We have now included an additional 'code and data availability' statement to make this clear to readers (see bottom of P42 in mark-up document) which reads:

"For persons interested in applying a similar signature-based LOA approach for model evaluation, we would encourage them to contact the authorship who are open to providing advice and sharing data and code where possible."

Ultimately, I find there is too much interpretation in the results chapter and suggest a clearer distinction between results and discussion

We purposely included some interpretation in the results section to help guide the reader through what are a complex set of results. However, we appreciate your preference for a complete separation of the results and discussion sections. **On re-reading the results, we've identified 10 portions of text with interpretation that can be reworded and moved into the discussion section: P23 L17, P25 L7, P29 L8, P30 L12, P31 L9, P32 L5, P32 L14, P32 L34, P35 L9, P35 L33. Accordingly, our revised manuscript will include these modifications.**

We have completely removed eight of the ten identified portions of text from the document as they made statements which were already included in the discussion section. The remaining two statements have been reworded and included in the discussion (P40 L1 and P40 L9 in the mark-up document).

Inconsistent order of citations P4L19

We used the Copernicus Latex template and we assumed these would be correctly ordered before publication, but **we are happy to re-order the citations ourselves if the editor requires.**

We have gone through the entire manuscript and re-ordered them manually (first by date, then by author).

Introduce, omit, or at least quote the abbreviation SEHR-ECHO P5L10

This will be addressed in the revised manuscript.

Removed as suggested (see P4 L27 in mark-up document).

Provide unit for slope (radians, I guess) Personally, I find slope angles in degree more convenient.

This will be addressed in the revised manuscript.

These have now been included (see P17 in mark-up document). Note, the units are in $m^3 s^{-1}$ per section of the flow duration curve. See in the attached figure of the FDC (white dotted line is the mean FDC and blue bars indicate uncertainty). The volume under the curve is equivalent to the integral between two exceedance limits on the x-axis. The slope, is the change in discharge per flow exceedance section. We've made this clear in the table caption in the revised manuscript.

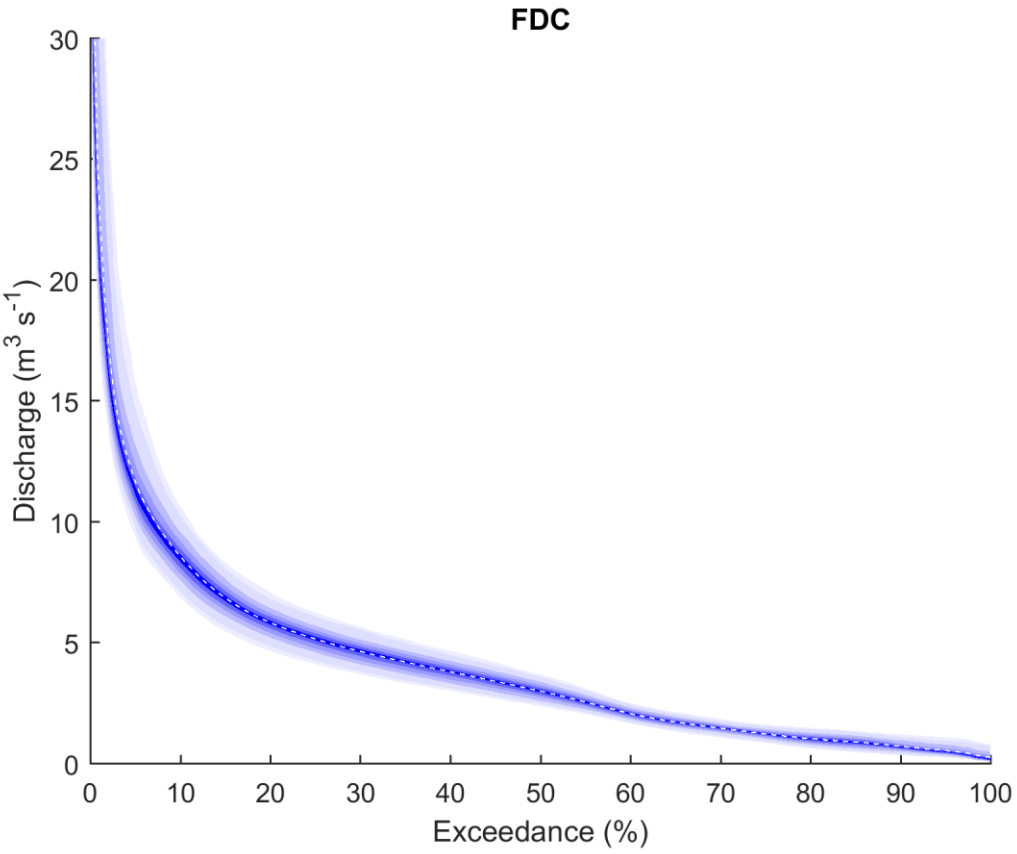

Fig. 9: Why are boxes dark gray where the score is 0?

This is rounding issue in the code used to generate these plots – i.e. those scores were not equal to zero, but very close. Indeed, this is also an issue for Figure 15 which we hadn't noticed. **These will be corrected in the revised manuscript.**

We have now modified these figures so that non-zero scores that round to zero for the plot are accompanied by a plus or negative sign to indicate the sign of the score, and to indicate that they are indeed non-zero and therefore unacceptable. Furthermore, additional text has been added to the figure captions to make clear to the reader that these scores are non-zero (see Figure 9, P27 and Figure 15, P33 in the mark-up document).

Fig. 13: What is '1e3 m3 '?

This is accepted standard exponential notation, which in this example is equivalent to 1×10^{-3} m^3. It was used to avoid excessively large numbers on the y-axis.

Anonymous Referee #3

This manuscript (ms) aims to presents "A limits of acceptability framework for model selection for glacio-hydrological melt and runoff modelling". For this purpose three model structures were calibrated and validated with ice volume change observations, satellite derived snow cover images and runoff data. The authors conclude that "it remains to be seen if the framework can be used", acknowledging that the results were "not necessarily more consistent across the full range of signatures". Further studies would be needed to investigate these conclusions.

I think the topic of this manuscript is highly relevant and important for hydrological modelling and therefore it should be tackled thoroughly and with care. In recent years many studies have compared model complexity and different observational data sets to investigate glacio-hydrological melt and runoff modelling. Accordingly, I would highly appreciate a concise and well-established framework of acceptability of glaciohydrological models. However, I have my major doubts about the present ms and if it delivers what the title promises. My concerns are the following: i) the authors conclude that the results were "not necessarily more consistent across the full range of signatures" and that "it remains to be seen if the framework can be used". Based on the presented results I would agree with these conclusions and will provide comments how this can be tacked in a more thorough way below. But I also would argue that a framework that cannot provide consistent results and has to be investigated further is not publishable. ii) The evaluation metrics (equation 9) are based on user defined signatures thresholds, rather than well-established efficiency criteria. This makes it impossible for the reader to assess the performance of the model. iii) Inter-annual average altitude dependent simulated and observed snow cover is presented (Fig 12) rather than daily time steps. This makes it impossible to assess how the model performs during years with enhanced snowfall and reduced snowfall. Averages of inter-annual performances dilute the performance, making it impossible for the reader to assess the real performance. iv) Same argument is valid for runoff: in my opinion daily time-steps should be compared for a calibration and validation period. v) In my opinion, a validation of the modelling results based on the presented "user defined" evaluation metrics is inadequate, because it is not objective (If other thresholds were selected the results might have been completely different, making the presented analysis subjective). vi) Many of the concerns addressed above have been analyzed and discussed in recent works, some of which are referenced in this ms but not taken adequately into account. In my opinion the authors fail to connect to previous works and point out why this framework is novel and how this ms fills an existing knowledge gap.

We echo the initial thoughts by Anonymous Referee #3 on the relevance of this kind of research and the need for these aspects to be tackled thoroughly and with care. This is certainly something we aimed to do in undertaking this work. The points above provide a useful overview of the referee's six

primary concerns and critiques of the manuscript, the majority of which are covered by subsequent, more specific comments below. As such, most of our detailed responses will be outlined later under these specific comments, but we feel it would also be useful to provide some initial thoughts before getting into these details.

On reading the full set of comments from Anonymous Referee #3, it is clear that one issue they have with the manuscript is that we are not using what could be termed 'traditional' means to evaluate the error, or skill, of the different models. For example, a number of critiques point to the fact that we are not evaluating 'daily' runoff or snow coverage patterns (although this is not entirely accurate, e.g. see Figures 13 and 16 for sub-daily runoff patterns) and that we are not using 'well-established efficiency criteria'. While we understand the referee's preference to use these kinds of criteria for evaluating model efficiency, we believe we set out clearly that this was not the premise of this manuscript. Its purpose, as outlined by the aims in the introduction (P4 L30), is to test a signature-based LOA approach which has been applied in hydrology (for good reason – see comments below on traditional efficiency criteria), and which we believe could also benefit the glaciology community given the similar problems of model equifinality (e.g. for melt models) demonstrated in the literature. Our thesis, which aligns with the recent literature (e.g. P4L9), is that we must go beyond the use of error criteria such as the RMSE or NSE (which describe an averaged or aggregated model skill, biased to high observational values) if we are going to diagnose deficiencies in model process representation, and support the identification of better model structural components, very importantly taking into account observational data uncertainty.

On reflection, we appreciate that the title of manuscript does not completely reflect the aims of the study and could be read as suggesting that we are presenting a definitive model selection framework when in actual fact the point of the study is to test the framework for identifying model deficiencies and selecting acceptable models. We feel at least some of the critique could stem from this initial framing of the study. **Indeed, as outlined in later comments, we suggest a small re-wording of the title in the revised manuscript that would help to frame this study properly**.

At this point, we also feel it would be worthwhile responding to the criticism about the subjectivity of the LOA (point v above) which is not explicitly raised in the remaining referee comments. To be clear, our use of the word objective is referring specifically to the definition of the acceptability thresholds and not the signatures themselves. **We perhaps have not made this completely clear, and if so will revise the text to ensure this is.** Of course, the choice of signatures will not be the same from study region to study region. For one, it will depend on what data are available. For example, we used river discharge signatures that characterise sub-daily behaviour of the system. Where only daily data are available, this won't be possible. The choice of signatures will also depend on which processes one is most interested in interrogating. For example, someone testing a distributed, physically based, snow-redistribution model who has high resolution snow depth measurements may derive signatures that thoroughly interrogate these detailed aspects of the model behaviour. For our study, we have derived a set of signatures that broadly characterise the glacio-hydrological behaviour of the basin based on the data we had available. However, these are by no means absolute. On the contrary, we would encourage others to experiment with different signatures to further interrogate model acceptability across different regions. Regardless of the choice of signatures, the LOA for a given signature should be defined from available information on observation data uncertainty. This definition of the acceptability criteria is the objectivity we refer to which is far more objective than some arbitrary efficiency criteria (see later responses).

To make this clear to the reader we have included some additional text at the end of the discussion section to emphasise that the choice of signatures is subject-specific (see P41 L31 of the mark-up document). It reads:

"Certainly, it's important to emphasise that these future applications need not adopt the same 33 signatures used in this study. On the contrary, the choice of signatures will always depend somewhat on the availability of data at a given study site as well as the complexity (e.g. spatio-temporal resolution) of the model(s) being interrogated. Indeed, future users should be encouraged to experiment with different signatures (where data permits) particularly if they wish to focus on other process representations within their GHM. Study sites with good observation data and understanding of data uncertainty would be ideal candidates for these future applications."

Additionally, we noted that in the original abstract, discussion and conclusions, we frequently referred to 'the LOA framework' which, on reflection we feel comes across as though we are proposing a definitive application of a LOA framework. As noted above, this is not the case, and accordingly we have changed a number of these instances to read 'a LOA framework' which we feel better reflects the fact we are presenting an application of a LOA framework (e.g. see P1 L18 in the mark-up document).

Finally, before proceeding to individual responses, we feel it is important to clarify the meaning of the two quotations from the manuscript which have been used in the above comments, but which we consider have been taken out of context and therefore used somewhat unfairly within the critique of the manuscript:

the authors conclude that the results were "not necessarily more consistent across the full range of signatures" and that "it remains to be seen if the framework can be used"

Regarding: "not necessarily more consistent across the full range of signatures"

This is referring to model complexity and has nothing to do with the suitability of the framework. To be clear, the full quotation from the manuscript reads:

"When evaluated against individual signatures, the more complex model formulations did improve model simulations in some cases. However, they were not necessarily more consistent across the full range of signatures, emphasising the need to exercise caution and properly evaluate if additional complexities are justified."

We believe this is an important point to emphasise, particularly in relation to representing water flow pathways through glaciated catchments. As we discuss in the introduction, linear reservoirs are widely used in glaciated regions to represent water flow pathways, but the form that these configurations should take is primarily down to a particular modellers own perceptions, experience in the field, model code affiliation etc…and rarely based on any formal exploration of model appropriateness. Note, we reference the paper of Hannah (2001) as one exception to this. One might expect (as we did), that given the known non-linearity in runoff behaviour in glaciated catchments because of the unique storage behaviour of snow and ice (e.g. see Jansson 2003), that a more complex (more non-linear) routing structure like ROR_3 in our study would better capture river flow regime over seasonal to sub-daily timescales – but we clearly show, using the range of river discharge signatures, that that isn't the case.

Regarding: "it remains to be seen if the framework can be used"

The full quotation from the manuscript reads:

"While all, but two, of the signatures demonstrated discrimination power, none of the 45,000 different model compositions tested in this study were able to capture them within their LOA simultaneously. Therefore, it remains to be seen if the framework can be used in this way, although we suggest that applications that go beyond examining the melt and runoff-routing structural uncertainties may prove more fruitful in obtaining a behavioural population of models."

So, to be clear we are reporting our finding for the second aim of the study as outlined in the introduction (P4 L29) which was to test the LOA framework for its ability to:

"ii) constrain a prior population of model structures and parameterisations down to a smaller population of acceptable models, indicating the framework's usefulness for reducing prediction uncertainty."

This is not to say it remains to be seen if the framework can be use as we know it can be used from applications in hydrology as outlined in the introduction (P4 L15). It is to say that finding a set of acceptable model structure/parameterisations could prove to be more difficult than we first thought. We should embrace this as something to work on (e.g. through the development of more suitable models) not something to dismiss.

Major concerns: 1) A framework has to work to be published. Accordingly, the framework should be able to reproduce adequately i) bi-annual glacier mass balances (accumulation and ablation phase) over several years, ii) daily snow cover ratio over several years to demonstrate that it works for snow intense and snow poor years. iii) daily runoff patterns.

A framework has to work to be published: We feel there might be some confusion as to what the purpose of this framework is and we expect this is linked to issues with the title. To be clear, it is not up to the framework to reproduce the observations. It is up to the models (including prescribed boundary conditions). The framework is there simply to indicate if a given model composition (i.e. structure and parameter set) is acceptable or not i.e. to determine if it is able to capture the signatures within their observation uncertainty. If none of the models are able to do this, this does not imply that the framework doesn't work, it implies that the models are not acceptable.

Accordingly, the framework should be able to reproduce adequately i) bi-annual glacier mass balances (accumulation and ablation phase) over several years: Again, it's important to reiterate here that it is not up to framework to reproduce available observation data. The framework's purpose is to tell the user whether a model is acceptable or not and what aspects of the system it succeeds or fails to capture (as defined by the signatures). In our case, we only have one set of melt season and one set of winter season ablation measurements (we'll discuss data availability below in a different comment). For other studies that have multiple years of ablation stake data, these could also be incorporated into the framework by adopting additional signatures as long as the uncertainty can be quantified, a point which we make in the discussions/conclusions.

ii) daily snow cover ratio over several years to demonstrate that it works for snow intense and snow poor years. iii) daily runoff patterns: Yes, as we noted above, we sympathise with the referee's preference for the more traditional analysis of model performance (e.g. through comparing time-series and using efficiency criteria). However, we must stress that this quite simply is not the point of this manuscript. We clearly outline that we are applying a different type of model evaluation technique: one that is based on signatures rather than time-series/efficiency criteria, which allows one to account for evaluation data uncertainty when making decisions about model appropriateness.

With regards to snow cover, it would have been possible to increase the number of snow signatures adopted in the study to incorporate snow intense/snow poor years. Doing this on a daily scale would have been difficult, mainly because good data are sparse (as noted in the methodology) and even when data do pass the QA, they often only cover parts of the catchment (presumably due to high relief and cloud cover). Accordingly, we deemed some degree of aggregation necessary to average out these discrepancies. In conjunction with this, we were also mindful that we already had 33 signatures to discuss, and given that it was clear none of the tested model compositions could capture the seasonal distribution of snowfall, we did not deem it necessary to further interrogate the models on extremes as the extra analysis, figures and text required simply was not worth it.

With regards to river discharge, we do relate model deficiencies identified in the framework to time-series of observed and simulated hourly river discharge for different flow regimes (Figures 13 and 16). In fact, given that river discharge signatures are well established in the literature (and because that's the thing that ultimately impacts downstream communities), we use a wide range of metrics over a range of timescales from monthly (monthly mean flows) to hourly (Integral time and peak flow hour – see Table 1).

2) If three model structures are used and they produce the same results, I would argue that the model structure is insensitive to the modelling approach. Nevertheless, the title claims to provide an acceptability framework for model structure. How can the framework say anything about model structure, it all the structures produce similar results?

Indeed, we found that all of the tested model structures were deemed unacceptable, i.e. they exhibit deficiencies that go beyond the known uncertainties of the evaluation data, and therefore none met the criteria to be 'selected'. On reflection we agree that we could have worded the title better so that it's clear we are testing the framework, rather than presenting it as a definitive model selection tool. Furthermore, we evaluated the framework for its ability to identify model deficiencies as well as for model selection which again, is not reflected in the title. As a reminder, and as set out at P4 L29, the aims of the study were to investigate the LOA framework utility for:

1) Diagnosing deficiencies in different model structures.
2) Constraining a prior population of model structures and parameterisations down to a smaller population of acceptable models.

Accordingly, we propose renaming the manuscript to better reflect both of these aims and emphasise the fact that we are applying the LOA framework for this purpose: **'Glacio-hydrological melt and runoff modelling: application of a limits of acceptability framework for model comparison and selection'.** We feel this is much more in line with the aims of the manuscript.

We have renamed the revised manuscript as described (see P1 of mark-up document).

3) Evaluation metrics should be consistent and comparable with previous literature. I understand that the chosen study site might lead to exceptionally low performance due to extreme weather patterns, but I would still use Nash-values for runoff, RMSE for glacier mass balances and a representative index for the snow cover area. This would enable a comparison of the model performance with previous works.

We're not completely sure if this is: i) a request to include these metrics on top what we have already presented; or ii) a critique of using signatures instead of these metrics (i.e. a critique of the framework). Accordingly, we will provide two responses here:

i)**We can include an additional table of these evaluation metrics in the appendix of the revised manuscript purely for comparison to other studies** (see later response for details of this). We chose not to do this, simply because the manuscript is already quite large, and these metrics have little to do with the framework. However, if the editor deems this necessary, we will include them.

ii) The use of signatures instead of well-established efficiency criteria underpins the methodology of the LOA framework. We do feel we've clearly explained and justified this in the introduction, but to be clear we'll summarise below exactly why signatures have been adopted in past applications instead of efficiency criteria:

1) What do efficiency criteria tell you about model efficiency? A key issue with using efficiency criteria to evaluate models is that they tell you very little about where the model is going wrong – they are an average of the model performance over a given observation dataset. We would ask, what does an NSE of 0.7 against observed river-flow time-series actually mean? What does it tell you about the model's ability to represent slow-release flow flows? What does it tell you about the model's ability to emulate the responsiveness (flashiness) of the system? Furthermore, all of these types of metrics are biased towards certain aspects of observation data. For example, it's been long recognised that the RMSE and NSE, both widely used to evaluate fit to river flow time-series, are biased towards fitting the highest flows (e.g. see Gupta et al 1998). MAPE on the other hand biases low flows. So using these not only provides very little information on where the model is working and where it is not, they also provide an overall biased indication of model efficiency.
2) Use of global efficiency criteria exacerbates the equifinality problem: Given 1), it's clear that evaluating models based on global metrics of model fit will give rise to multiple model structures and parameterisations that produce similar overall model fit leading to higher prediction uncertainty (e.g. see Gupta et al 2008 referenced in manuscript).
3) How do you define acceptability criteria of a model efficiency metric? The advantage of using signatures as evaluation metrics, is not only that they allow you to analyse specific aspects of a model's behaviour, but that quantitative LOA can be defined around these based on information about observation uncertainty providing the ability to evaluate model performance within the uncertainty (see P4 L9) of the evaluation metrics (especially important in mountain regions where observation data are often riddled with uncertainty). Acceptability criteria based on efficiency metrics have been used widely in the past. E.g. for river flow predictions, NSE > 0.6 is frequently used. However, this number is completely subjective (some use NSE > 0.5). What's more, given 1), one has no idea about what kind of model errors are being introduced to the predictions based on these criteria.

We appreciate that using signatures as the basis for model evaluation may seem controversial, but as we clearly set out in the introduction, it is something that has been done in the hydrology community and something that we feel is completely relevant to the glaciology community as well.

4) In my opinion, the ms can be significantly shortened, made more concise and the literature should be selected more carefully.

See Reponses to specific comments about text length and literature.

Specific recommendations: Title: I find the title confusing and misleading. Please ´ provide a more focused title

We agree that the title could be more focussed **and we have proposed modifying it for the revised manuscript.**

Abstract: do models underpin the understanding of future climate change? I would argue that only the analysis of the results obtained with models can do that (it's a detail, but important, I recommend that the entire ms is revised and such flaw statements are corrected)

We agree, it's a technicality, but an important one. Thank you for pointing this out. **We will change this and any similar statements so that it reflects the fact that it is the analysis of the projections from models that underpin our understanding.**

We identified two uses of this statement and have changed both to read "Glacio-hydrological models (GHMs) allow us to develop an understanding of…" (see P1 L1 and P1 L20 in mark-up document).

Introduction: in my opinion this chapter can be significantly shortened to a third of its current length. But I also think that it should be more targeted and the cited literature should be more focused.

We appreciate the introduction is longer than average, although we really can't see how we could remove two-thirds of it. To do so would surely mean removing key elements of the literature review. If this is the referee's recommendation, could we ask them to be more specific about what aspects of the introduction they believe should be removed?

We are also unsure as to what exactly 'more targeted' means. We presume this could be related to point vi from their initial comments which reads: "Many of the concerns addressed above have been analysed and discussed in recent works, some of which are referenced in this ms but not taken adequately into account. In my opinion the authors fail to connect to previous works and point out why this framework is novel and how this ms fills an existing knowledge gap."

Again, we are struggling to understand how to address this. We feel we've been explicit about what has and hasn't been done before. If the referee thinks we have missed something, could they please specify what this is? Which referenced material has not been taken into account properly? Which aspect of the novelty is not made clear? We do hope they can provide this information so that we can revise the introduction if required.

Referencing: in my opinion 3 references are sufficient for one statement; in this ms up to 17 references are used of a single statement. This is misleading and not helpful for the reader. I think every references should be carefully chosen and only the most relevant references should be used (this would keep the reader focused on the topic of the ms).

On reflection, we agree with the referee that the 17 citations (P19 L18) could be revised down to a smaller number. All of these were included simply because all were used to help determine which signatures to use in this study. However, several of the more recent publications include the majority of these and as such **we will edit this for the revised manuscript.**

We have revised this down to four publications which include the majority of the signatures used in this study (see P21 L5 in mark-up document).

We also acknowledge that the four citations used on P9 L7 could be revised down to the Huss *et al*. (2010) citation only, given that it is this study that best demonstrates that the delta-h parameterisation exhibits behaviour comparable to complex 3-D finite-element ice flow models. The other cited studies demonstrate additional applications of the approach rather than specifically testing it against complex ice-flow models. Accordingly, **we will edit this for the revised manuscript.**

This has been edited as described in the revised manuscript (P9 L17 in mark-up document).

While we appreciate the referee's preference for including a maximum of three citations, we do feel that in some cases it is useful to include more to provide the reader with some appreciation of the breadth of research that has been done a particular subject. For example, we use five references on P1, L20: 'Computational GHMs underpin our current understanding of how future climate change will affect river flow regimes in glaciated watersheds (Ragettli et al., 2016; Singh et al., 2016; Teutschbein et al., 2015; Lutz et al., 2014; Radic and Hock, 2014).' We chose these five studies because they cover applications of GHMs across a number of different regions (e.g. Lutz et al., 2014 in Asia vs Teutschbein et al., 2015 in Northern Europe) and at a number of different scales (e.g. Radic and Hock, 2014 global scale vs Ragettli et al., 2016 catchment scale).

Similarly, on P1 L22, we use six references: 'A variety of GHM codes exist (e.g. Bergström, 1997; Ciarapica and Todini, 2002; Schulla, 2015; Huss et al., 2008b; Boscarello et al., 2014; Schaefli et al., 2014), each of which include…' Here we chose a handful of the most well established GHM codes and some that are less well established but which demonstrate different levels of model complexity (e.g. the semi-distributed HBV model of Bergström, 1997 vs the physically-based TOPKAPI model of Ciarapica and Todini, 2002). WaSiM is another widely used model code (Schulla, 2015) while the model of Huss et al., 2008 has also been widely applied within a glacio-hydrological context. The model of Schaefli et al. (2014) is a recent addition to available GHMs and is unique in that it was originally developed for eco-hydrological modelling, but has been applied for glacio-hydrological modelling as well.

Also, I would recommend referencing the first publication that has investigated a topic, rather than referencing newer articles who have just build on previous works. Referencing is a tedious job, involves a lot of reading but should be taken seriously.

We understand the referee's preference for citing the oldest publications on a given topic. We spent considerable time selecting the most relevant citations to include in the introduction (although we appreciate that on the two occasions noted above we could have been more prudent). Our overwhelming feeling was that for the topics covered, the most relevant literature was generally also relatively recent. This was not surprising, given the relative novelty of signature-based applications of the 'limits of acceptability' methodology and the clear need for new model interrogation/selection strategies within the field of glaciology. However, we would be open to suggestions of specific citations that the referee feels have been missed and would add important additional points to the introduction.

I would tend to disagree that this is the "first kind to apply a signature-based ´ limits of acceptability (LOA) framework". In my opinion numerous studies have done this, simply in a different manner.

We agree that this is not the first of its kind to apply a signature-based limits of acceptability (LOA) framework, but do not claim it to be. We consider that we clearly reviewed past studies that have applied a signature-based limits of acceptability framework (P4 L11). Rather, we state that: "This study is the first of its kind to apply a signature-based LOA framework for a multi-GHM-structure evaluation." i.e. we have taken something that has been used in hydrology and applied it in a glacio-hydrological context. We feel this is quite clear.

Figure 1: Why are meteo-data presented as an inlet in the map of the study site? Please stick to a coherent structure and provide all figures in a standard scientific style.

The meteo-data are presented as an *inset* because it is part of the catchment description. We also use insets in Figure 10. We do feel this is in keeping with standard scientific style and makes for a more compact figure so we would prefer to keep it as it is.

Observational data: in my opinion this chapter can also be significantly shortened. None of the data were collected by the authors, so the methods how they were collected can be found in the relevant literature.

We agree there is scope for shortening this section. **We will do so in the revised manuscript.**

We have now removed a significant portion of the text describing the climate data in section 2.2.1. In particular, we have removed the text describing the rain-gauge apparatus (P7 L14 in mark-up document) which on reflection, we feel was providing too much detail. Also, we've removed the description of how the ICRA precipitation data was produced (P7 L24 in mark-up document) given that this can be found in the cited literature.

Additionally, we have also removed the first sentence describing the snow coverage data in section 2.2.3 which only serves to reiterate a point covered in the introduction (P8 L11 in mark-up document).

Glacio-hydrological model: has the model never been published before? Does it have a name? Why not use a model that is well established? Then the description of the model could be shortened and the focus could be on the framework.

No the model code has not been published and it doesn't have a name. The reason we didn't use someone else's model code was essentially because there wasn't one out there that met all of our five requirements to ensure we could implement the LOA framework with available observation data at this study catchment:

- Inclusion of multiple TIM structures (calculated on distributed grid)
- Mass-conserving dynamic glacier evolution model
- Inclusion of multiple linear reservoir runoff-routing routines with dynamic HRUs
- Dynamic temperature lapse rates with on-ice temperature correction.
- Ability to run on HPC

While many codes include one or several of these, none (to our knowledge) include all of them. One option of course, would have been to use multiple model codes, but a very important problem with this (ignoring time requirements) stems from the fact that when it comes to interpreting differences between model acceptability, it is very difficult to determine what aspect of the model brought this about (e.g. was it difference in resolution, different in climate interpolation strategy, difference in time stepping etc.) Other considerations which led us to our own code was:

- Past experience of modifying open source code and potential to introduce errors.
- Ease of extracting required outputs for signature comparison.
- Runtime (many hydrological codes include extra processes that increase runtime).

Driving climate data: The Icelandic Met Office (IMO) provides gridded reanalysis data (P, T and I). They are continuously updating their gridded data sets. Accordingly, I would recommend using the official data set of the IMO, rather than reproducing something that has been investigated for many years by the IMO.

We maintained close contact with the IMO to ensure that we were aware of their most up-to-date climate datasets. In fact, we used their latest ICRA gridded reanalysis data for precipitation (see methodology section). Therefore, we are not entirely sure what you mean by 'official' dataset. Consequently, on receiving this comment we contacted the IMO again who assured us the ICRA dataset is the best available gridded precipitation product currently available (better than say the

Chrochet *et al.* (2007) data which only extends to 2007). For reassurance of this superiority, please see section 6 of the Nawri 2017 referenced in the manuscript.

While we could also have used reanalysis data for temperature and incident solar radiation, we decided to use available observation data taken by the AWS in the catchment and from any nearby IMO weather stations primarily because the reanalysis data is constrained by a sparse set of met stations that are almost entirely located at low altitudes (see Chrochet 2011 for distribution of weather stations that measure temperature in Iceland – I is even lower density). Most importantly, the catchment AWS provide vital information on temperature lapse rates and their variability which are highly catchment-specific and important for simulating ice and snow melt e.g. see Gardner and Sharp (2009) for one example showing this. For these reasons, we preference observation data obtained within the study catchment and from nearby weather stations as opposed to reanalysis data constrained on a sparse network of weather stations.

Figure 3: It is impossible to assess how well the correction worked based on the presented data; see my previous comment. I recommend computing some relevant statistical values (see some of the cited literature).

Yes, on reflection we completely agree that we have not provided adequate information on potential weaknesses of the precipitation data as also noted by referee #2, although we did include R^2 values within the text. Therefore**, for the revised manuscript we will include a table of statistics for the precipitation data include seasonal means, standard deviation, coefficient of variation and skewness for the observed and simulated data.**

The described table of statistics has now been included in the mark-up document (see table 1, P12 in mark-up document). We have also included additional text in section 2.4.1 (P13, L21 onwards in mark-up document) which describes the relative strengths and weaknesses of the bias-correction procedure and the precipitation data. Note, we then draw upon these in the discussion section in response to the comments from referee #2.

Fig 4,5,67: please see my concern regarding aggregated monthly ´ or inter-annual comparison of observed and simulated runoff and snow cover; What is the use of modelling daily time steps if only mean monthly results are evaluated? If the objective is only the get monthly means correctly than I recommend using a monthly time step in the modelling framework.

Yes we have somewhat covered this in our comments above, but to clarify, the model runs on an hourly time step and not a daily time step (see methodology section). Also, note that the river discharge signatures characterise model behaviour from monthly to hourly timescales (see table 1) and so are completely dependent on us running the model at a hourly resolution. We hope that makes it clear why we're running it on an hourly time step.

Fig 8: see my comment regarding the evaluation metrics

Discussed in previous responses.

Fig 9: I recommend making a table rather than a figure

We appreciate your preference for using a table rather than a diagram, however, having experimented with using a table previously we found the diagram to be much clearer to the reader, particularly because it allows the inclusion of two numerical values in each cell (one large colour-coded value and one smaller value confined to the top corner of the cell). We found including both

of these in a table (similar in size to table 1) became cluttered and was not clear to the reader. We thank you for the recommendation, but we would prefer to keep these results in diagram format.

Fig ´ 10: why select a case study which has only two glacier volume observations and 23 year of data gap between the two observations? There are numerous case studies which have bi-annual glacier mass balances, which would be a lot more valuable to establish a framework of acceptability for glacio-hydrological models; If the authors want to establish a transparent framework, I recommend selecting a case study with enough observational data to test the framework thoroughly.

We appreciate that there are other case studies with more observation data and there are also countless case studies with less observation data than we have. However, the amount of observation data by no means precludes the use of the LOA framework any more than it precludes the use of more traditional methods of model evaluation. That being said, we do agree that we would like to see the method applied to sites with even more observation data and better constrained driving climate data and we do make this clear in the discussion. Nevertheless, given the general pattern of data scarcity of mountainous regions, we feel this study region is as good as any for testing such a framework.

Fig 11: in my opinion these results suggest that the framework does not work; none of the simulations are acceptable, even according to the standards set by the authors.

We have taken care to address this in our previous responses.

Fig 12: this figure ´ illustrated the discrepancy between observed and simulated snow cover: In my opinion a acceptability framework should be able to identify if a model works during snow intensive and snow poor years. Here we only see that it never really works.

Again, we have taken care to address this in detail above. We would like to reiterate that none of the tested models were deemed acceptable across *all of the signatures*. This does not mean the framework did not work.

Final remark: I do think that this study can become an important contribution to hydrological modelling. However, all of the comments above would need to be accounted for and/or addressed. I would like to encourage the authors to have a thorough discussion on the purpose of this study. Furthermore, I encourage the authors to read some of the literature that is already referenced; many of the concerns addressed above have already been investigated and solutions have been presented. I would like to wish the authors lots of success and good luck with further research on such a framework.

We appreciate this extremely thorough set of referee comments.

In addition to all of the revisions noted above, we also made an additional minor revision after re-reading the manuscript (P4 L20 in the mark-up document). We noted that the sentence, “Such an approach was first proposed by Beven (2006), where observation data uncertainty could be used to define quantitative `limits of acceptability' (LOA) around model evaluation metrics” was making the same point as P3 L23. Accordingly, this has been removed and re-worded.

References:

Crochet, P., Jóhannesson, T., Jónsson, T., Sigurðsson, O., Björnsson, H., Pálsson, F., & Barstad, I. (2007). Estimating the Spatial Distribution of Precipitation in Iceland Using a Linear Model of

Orographic Precipitation. *Journal of Hydrometeorology*, *8*(6), 1285–1306. https://doi.org/10.1175/2007JHM795.1

Crochet, P., & Jóhannesson, T. (2011). A data set of gridded daily temperature in Iceland, 1949-2010. *Jökull*, *61*, 1–17.

Gardner, A. S., & Sharp, M. (2009). Sensitivity of net mass-balance estimates to near-surface temperature lapse rates when employing the degree-day method to estimate glacier melt. *Annals of Glaciology*, *50*(50), 80–86. https://doi.org/10.3189/172756409787769663

Glacio-hydrological melt and runoff modelling: application of a limits of acceptability framework for model comparison and selection

Jonathan D Mackay[1,2], Nicholas E Barrand[1], David M Hannah[1], Stefan Krause[1], Christopher R Jackson[2], Jez Everest[3], and Guðfinna Aðalgeirsdóttir[4]

[1]School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
[2]British Geological Survey, Environmental Science Centre, Keyworth, Nottingham, NG12 5GG, UK
[3]British Geological Survey, Lyell Centre, Research Avenue South, Edinburgh, EH14 4AS, UK
[4]Institute of Earth Sciences, University of Iceland, 101 Reykjavík, Iceland

Correspondence to: Jonathan D Mackay (joncka@bgs.ac.uk)

Abstract. Glacio-hydrological models (GHMs) ~~underpin our understanding~~ allow us to develop an understanding of how future climate change will affect river flow regimes in glaciated watersheds. A variety of simplified GHM structures and parameterisations exist, yet the performance of these are rarely quantified at the process-level or with metrics beyond global summary statistics. A fuller understanding of the deficiencies in competing model structures and parameterisations and the ability of models to simulate physical processes require performance metrics utilising the full range of uncertainty information within input observations. Here, the glacio-hydrological characteristics of the Virkisá river basin in southern Iceland are characterised using 33 'signatures' derived from observations of ice melt, snow coverage and river discharge. The uncertainty of each set of observations are harnessed to define 'limits of acceptability' (LOA), a set of criteria used to objectively evaluate the acceptability of different GHM structures and parameterisations. This framework is used to compare and diagnose deficiencies in three melt and three runoff-routing model structures. Increased model complexity is shown to improve acceptability when evaluated against specific signatures, but does not always result in better consistency across all signatures, emphasising the difficulty in appropriate model selection and the need for multi-model prediction approaches to account for model selection uncertainty. Melt and runoff-routing structures demonstrate a hierarchy of influence on river discharge signatures with melt model structure having the most influence on discharge hydrograph seasonality and runoff-routing structure on shorter-timescale discharge events. None of the tested GHM structural configurations returned acceptable simulations across the full population of signatures. The framework outlined here provides a comprehensive and rigorous assessment tool for evaluating the acceptability of different GHM process hypotheses. Future melt and runoff model forecasts should seek to diagnose structural model deficiencies and evaluate diagnostic signatures of system behaviour using ~~the~~ a LOA framework.

1 Introduction

Computational GHMs ~~underpin our current~~ allow us to develop an understanding of how future climate change will affect river flow regimes in glaciated watersheds ~~(Ragettli et al., 2016; Singh et al., 2016; Teutschbein et al., 2015; Lutz et al., 2014; Radić and Hock,~~

(Lutz et al., 2014; Radić and Hock, 2014; Teutschbein et al., 2015; Ragettli et al., 2016; Singh et al., 2016). A variety of GHM codes exist ~~(e.g. Bergström, 1997; Ciarapica and Todini, 2002; Schulla, 2015; Huss et al., 2008b; Boscarello et al., 2014; Schaefli et al., 20~~ (e.g. Bergström, 1997; Ciarapica and Todini, 2002; Huss et al., 2008b; Boscarello et al., 2014; Schaefli et al., 2014; Schulla, 2015), each of which include a number of model components that represent two broad groups of processes: i) glaciological mass balance: the accumulation and ablation of snow and ice; and ii) hydrological water balance: the storage and release of melt and rainfall through snow, ice, overland and the subsurface. The exact form that these model components should take, both in terms of their governing equations (structure) and numerical constants (parameterisation) is not known. Physically-based models which solve equations derived from first principles, typically over a distributed grid, are our closest approximation of the 'true' structure. However, limited parameterisation data and computer resources often preclude the use of such complex models, particularly in remote mountainous regions where data are scarce and where the inclusion of extra complexity does not guarantee better predictions (e.g. Gabbi et al., 2014).

Simplified process models offer an alternative. They are faster to run and employ fewer parameters that are typically calibrated to available observation data. They are based on, but do not necessarily adhere to, physical laws and as such their mathematical structure is somewhat unconstrained and may be biased towards a particular scientist's own perceptions and understanding of environmental processes. This has led to the development of a variety of competing model structures which purport to simulate the same process, but which have been derived from different process hypotheses. For example, a number of simplified 'index' model structures of snow and ice melt exist. The classical temperature index model (TIM) simulates melt as a linear piecewise function of temperature only (Braithwaite, 1995), a hypothesis that can be justified because of the influence temperature has on the total energy balance of ice and snow, particularly in temperate climates ~~(Aðalgeirsdóttir et al., 2011; Guðmundsson et al., 2009; Ohmura, 2001)~~(Ohmura, 2001; Guðmundsson et al., 2009; Aðalgeirsdóttir et al., . So-called 'enhanced' TIM structures have also been proposed which include added levels of complexity with the purpose of providing more accurate estimates of melt. These have accounted for perturbations in melt caused by topographic shading (Hock, 1999), surface albedo characteristics ~~(Pellicciotti et al., 2005; Oerlemans, 2001)~~ (Oerlemans, 2001; Pellicciotti et al., 2005) and debris cover (Carenzo et al., 2016).

Similarly a number of simplified representations of processes governing the hydrological water balance have been used in GHMs. Arguably, the equations that govern the routing (transport) of runoff are most important in relation to river flow predictions in glaciated river basins, as storage characteristics of ice and snow strongly influence river flow regimes over a range of time-scales (Jansson et al., 2003). The concept of linear reservoirs is the most widely adopted simplified approach for runoff-routing in glaciated basins (Zhang et al., 2015; Hanzer et al., 2016; Gao et al., 2017). A linear reservoir lumps all of the interacting, non-linear and non-stationary components of water transmission within a pre-defined area (e.g. a watershed) into a single 'leaky bucket'. Despite its simplicity, the linear reservoir has shown to be remarkably versatile at capturing the storage-discharge characteristics of glaciated river basins around the world (Hock and Jansson, 2005; de Woul et al., 2006; Farinotti et al., 2012). This is partly because the concept lends itself to structural modifications that can represent different glacio-hydrological systems. Hanzer et al. (2016) hypothesised that the snow pack, firn layer, glacier ice and the region free from ice all exhibit unique runoff-discharge responses and advocate the use of four linear reservoirs in parallel to distinguish

between these units. However, simpler structural configurations using only two linear reservoirs in parallel to route meltwater through the snowpack and ice separately (Hannah and Gurnell, 2001) or even a single linear reservoir to route all rainfall and melt runoff simultaneously (Boscarello et al., 2014) can accurately reproduce river discharge time-series.

The availability of multiple, presumably plausible, simplified model structures presents somewhat of a dilemma to glaciologists and hydrologists as they are left with some uncertainty about how processes should be represented in their models. For the purpose of river discharge predictions, this problem is particularly pertinent as there are competing structures for two fundamental controls on these predictions: snow and ice melt and runoff-routing. One approach to mitigate this is to determine the 'optimum' structure that best captures the observation data. Structural optimisation of simplified runoff-routing routines has largely been ignored in glacio-hydrological contexts (see Hannah and Gurnell, 2001, for one notable exception), but more studies have sought to optimise and compare simplified models of melt. Gabbi et al. (2014) applied four different TIMs to Rhonegletscher, Switzerland. They found that all achieved a similar goodness-of-fit to six years of ablation stake data, but that the inclusion of a solar radiation term provided the most accurate predictions of multi-decadal measurements of ice volume change. Irvine-Fynn et al. (2014) applied six different TIMs to the High-Arctic Midtre Lovénbreen glacier but only found minor improvements at capturing seasonal ablation stake data when various levels of complexity were introduced to the classical (temperature-only) TIM. More recently, a comparison of four TIMs applied to four glaciers in the French Alps by Reveillet et al. (2017) found no clear evidence that using an enhanced TIM over the classical temperature-only approach provided better simulations when compared to a 17-year dataset of ablation stake measurements. Mosier et al. (2016) used a multi-criterion evaluation approach to compare the performance of different conceptual melt model structures. They compared seven competing melt model structures in two glaciated catchments in Alaska to ablation stake, river discharge and remotely-sensed snow coverage data. They found that no single model was best across all of the observation datasets, but the inclusion of a snow cold content representation consistently produced the best goodness-of-fit scores over the evaluation data.

Clearly, while some studies have provided useful insight into the comparative behaviour between competing conceptual process hypotheses (particularly for melt), none provide any definitive reasoning for adopting (or not) a particular model structure. Of course, discriminating between competing model structures in this way is made difficult by the fact that observation data used to drive and evaluate models are uncertain and therefore ~~,~~ we cannot be sure whether model deficiencies represent inadequacies in the model or the data (Beven, 2016). Beven (2006) argues that because of this uncertainty and because of the fact that all models are by definition ~~,~~ imperfect, no one optimum model structure (or parameterisation) exists. Instead, there is an equifinality of 'behavioural' models that make predictions within some pre-defined acceptability bounds around the observation data that take account of the various sources of modelling uncertainty. Indeed, parameter equifinality is a well recognised phenomenon in conceptual models of snow and ice melt ~~(Gabbi et al., 2014; Jost et al., 2012; Finger et al., 2015; Pellicciotti et al., 2012; Revei~~ (Jost et al., 2012; Pellicciotti et al., 2012; Gabbi et al., 2014; Finger et al., 2015; Reveillet et al., 2017). If we accept this, then a priority within the glacio-hydrological modelling community should be to establish frameworks that allow us to robustly evaluate model appropriateness and distinguish between behavioural (acceptable) and non-behavioural (unacceptable) structures and parameterisations. Constraining the range of acceptable models is particularly important for glacio-hydrological modelling

as it has been shown that model uncertainty can lead to high uncertainty in 21st century predictions of river flows in glaciated basins (Huss et al., 2014).

One potential source for inspiration is the hydrological rainfall-runoff modelling community. Their heavy reliance on an ever-expanding choice of conceptual hydrological process models to make river flow predictions prompted Gupta et al. (2008) to discuss the need for a better framework in which to discriminate between these competing process hypotheses. They focussed on the evaluation metrics and noted that there was an over-reliance on metrics that quantify the average performance of a model (e.g root mean squared error and Nash-Sutcliffe efficiency) which reduce information held in observation data down to a single summary statistic. They argue for a multi-criterion, 'diagnostic' approach where more of the relevant information from observation data is extracted so that inadequacies in model structures and parameterisations can be better identified. Rye et al. (2012) applied such an approach to optimise a distributed surface mass balance model of two glaciers in Svalbard. They used ablation stake data to define three different features of the observations including mass balance at the stake locations, long term mass balance trend and mass balance gradient. Using a multi-objective optimisation procedure, they identified structural inadequacies relating to how the mass balance gradient was simulated.

~~Indeed, hydrologists~~ Hydrologists are now moving away from traditional metrics of model performance in favour of more diagnostic 'signatures' of hydrological behaviour. These have typically been derived from river flow time-series and may be as simple as the mean flow (an indicator of water balance) or they can be used to characterise the distribution (e.g. flow percentiles) and the timing (e.g. autocorrelation) of flows. They have shown to have more discrimination power than traditional error metrics ~~(Hrachowitz et al., 2014; Shafii and Tolson, 2015; Euser et al., 2013; Schaefli, 2016)~~ (Euser et al., 2013; Hrachowitz et al., 2014; Shafii an and, importantly, it is also possible to take account of their information content (i.e. their uncertainty) so that decisions about model appropriateness can be made within the uncertainties of observation data used to evaluate the model. ~~Such an approach was first proposed by Beven (2006), where~~ Here, observation data uncertainty ~~could~~ can be used to define quantitative 'limits of acceptability' (LOA) around ~~model evaluation metrics. Using signatures as the basis for model evaluation, different model~~ each signature. Different model structures and parameterisations can then be systematically evaluated for their ability to capture the signatures within their LOA, allowing the modeller to objectively diagnose model deficiencies and make decisions about model appropriateness. The LOA framework has been used to constrain the parameters of a distributed hydrological model for flood prediction (Blazkova and Beven, 2009), evaluate the appropriateness of different hydrological model structures across contrasting geological settings (Coxon et al., 2014) and, most recently, to diagnose deficiencies in ~~the SEHR-ECHO GHM~~ a hydrological model based on its ability to capture a range of river discharge signatures for an Alpine catchment (Schaefli, 2016).

A signature-based approach within a LOA framework could also be used to compare and diagnose deficiencies in different simplified melt and runoff-routing model structures and parameterisations employed in GHMs. For this purpose, signatures need not be derived just from river discharge data, but should also be taken from other observation sources such as ice melt and snow coverage as these have shown to be useful for evaluating the consistency of GHMs across different aspects of glacio-hydrological systems ~~(Finger et al., 2015; Hanzer et al., 2016; Mayr et al., 2013; Finger et al., 2011)~~(Finger et al., 2011; Mayr et al., 2013; . By doing so, this framework could facilitate better predictions of river flow regime changes in glaciated river basins; firstly by

helping to diagnose deficiencies in GHM structures that require improvement, and secondly, by objectively selecting the range of acceptable model structures and parameterisations so that prediction uncertainty can be better constrained.

This study is the first of its kind to apply a signature-based LOA framework for a multi-GHM-structure evaluation. The framework is used to evaluate three different melt model structures and three different runoff-routing model structures with the aim of investigating its utility for: i) diagnosing deficiencies in the different model structures, indicating the framework's usefulness for aiding future improvement of simplified process models; and ii) constraining a prior population of model structures and parameterisations down to a smaller population of acceptable models, indicating the framework's usefulness for reducing prediction uncertainty. To do this, the models were applied to the glaciated Virkisá river basin in southern Iceland where observation data were used to derive 33 signatures of ice melt, snow coverage and river discharge against which the models were calibrated and evaluated. LOA were defined around each signature so that acceptable and unacceptable model structures and parameterisations could be defined. The results were first used to evaluate the capacity of the different signatures for discriminating between acceptable and unacceptable model structures and parameterisations when used individually. They were then used to compare the acceptability of the different melt and runoff-routing model structures across the range of signatures.

2 Methodology

2.1 Study site

The Virkisá river basin covers an area of 22 km^2 on the western side of the ice-capped Öræfajökull stratovolcano in south-east Iceland (Fig. 1). It rises from near sea level to the west, where it is bounded by steep cliffs, up to an ice-filled caldera at the summit of Öræfajökull (∼2000 m asl), the edge of which forms the basin's uppermost boundary. The basin forms a major drainage channel for accumulated ice which flows in a south-westerly direction down the steeply-sloped Öræfajökull (average slope of 0.25). It flows along two distinct glacier arms (Virkisjökull and Falljökull, hereafter referred to as Virkisjökull) around a bedrock ridge before meeting again at the terminus (∼ 150 m asl). Virkisjökull has a high mass balance gradient with a net annual accumulation of more than 7 m w.e. y^{-1} at the summit (Guðmundsson, 2000) and net annual ice melt of more than 8 m w.e. yr^{-1} in the main ablation zone (Flett, 2016). It has been in a phase of retreat since 1990 due to warming of the climate over this period (Hannesdóttir et al., 2015). Since 2005 the rate of retreat has accelerated to >30 m yr^{-1} as a result of the detachment ~~on~~ of the ice front from the active part of the glacier, resulting in rapid down wasting ~~(Bradwell et al., 2013; IGS, 2017; Phillips et al., 2014)~~(Bradwell et al., 2013; Phillips et al., 2014; IGS, 2017). This recent rapid retreat has resulted in the formation of a small proglacial lake at the terminus which forms the headwater of the Virkisá river. The Virkisá river flows south-westerly, firstly through a 800 m bedrock-controlled section flanked on either side by push moraines from previous glacial advances. From here it continues to flow over an extensive and gently sloping sandur floodplain. The steep-sided valley walls and the relatively recent glacial maximum at the end of the Little Ice Age circa 1890 mean that there is limited soil development in and around the Virkisá river basin. Where thin soils have developed, vegetation is dominated by mosses, sparse grass and shrubs such as dwarf willow and birch.

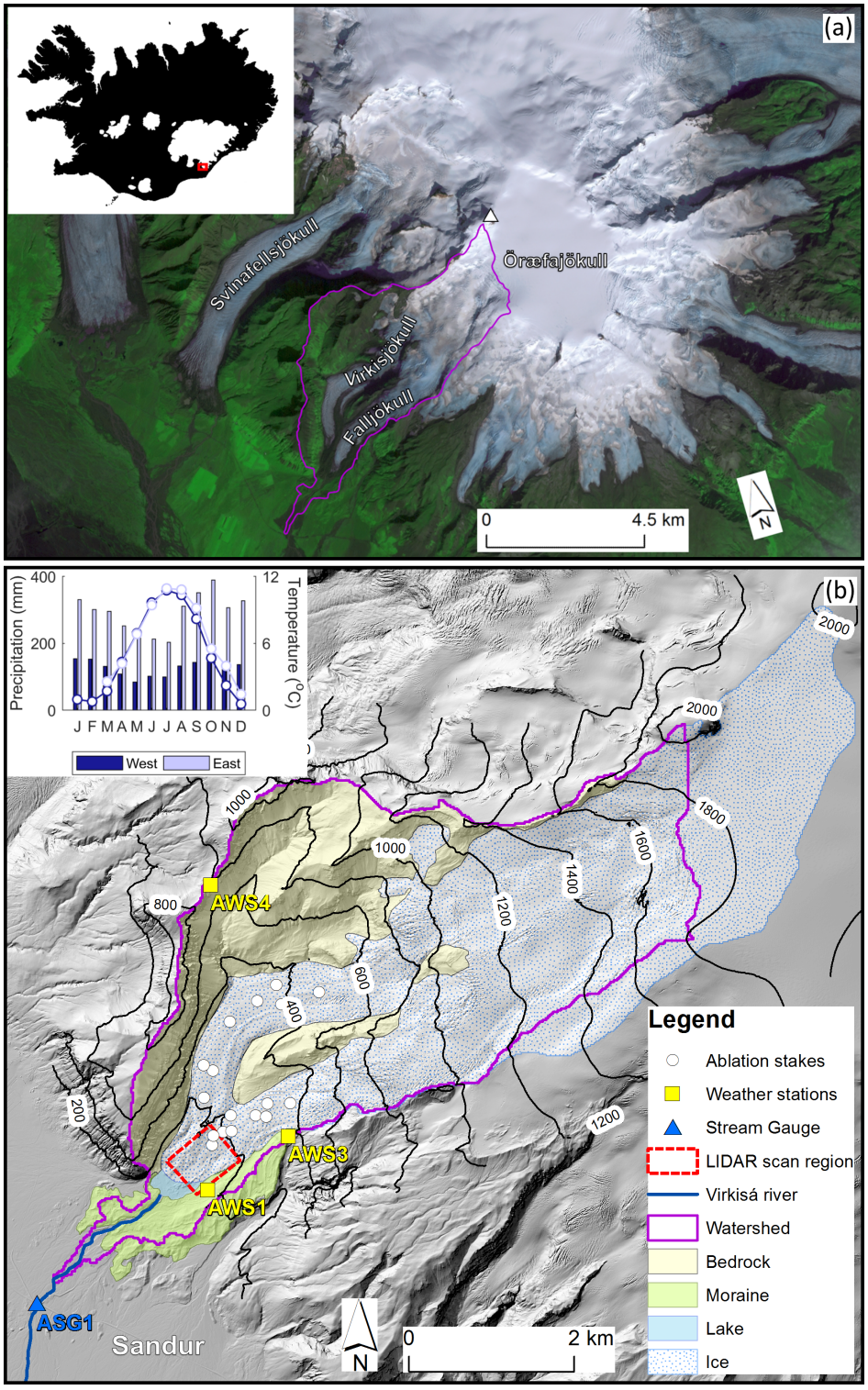

Figure 1. Location of Virkisá river basin on Öræfajökull (top) and detailed topographical map of basin including major land surface types and observation data with inset showing mean monthly climate (bottom).

Long-term meteorological records from two weather stations operated by the Icelandic Meteorological Office 10 km east and west of the study site show that the region experiences a maritime climate characterised by cool summers ($\sim$ 10 ^{o}C on average) and mild winters ($\sim$ 1 ^{o}C on average) with year-round precipitation (see inset in Fig. 1 bottom). The prevailing north-easterly wind and orographic lift over Öræfajökull induces a strong lateral precipitation gradient where more than two-times the precipitation falls to the east (3500 mm yr^{-1}) of the river basin than to the west (1500 mm yr^{-1}). Near-surface air temperature is mainly controlled by the altitudinal variations over Öræfajökull where the average temperature lapse rate is -0.44 ^{o}C 100 m^{-1} (Flett, 2016).

2.2 Observation data

2.2.1 Climate

Several different sources of climate data are available for the study site. Measurements within the catchment are available from three automatic weather stations (AWS) installed by the British Geological Survey (BGS) between 2009 and 2011 as part of their investigation into the retreat of the Virkisjökull glacier. These are situated at 156, 444 and 805 m asl (Fig. 1) and they measure temperature, air pressure, humidity, wind speed and rainfall every 15 minutes. The lowest weather station (AWS1) is also equipped with a cosine-corrected pyranometer which measures incident shortwave radiation. ~~Reliable, continuous time-series are available for the majority of weather variables, however, the continuity of the rainfall records are dependent on snowfall. More specifically, AWS1 is fitted with a tipping bucket rain gauge while AWS3 and AWS4 are fitted with Vaisala RAINCAP® acoustic sensors that measure the impact of individual raindrops on a steel plate. Neither of these apparatus are~~ None of the weather stations are designed to measure snowfall ~~. Furthermore, the presence of snowfall and /or the freezing of residual rainfall on the devices may induce erroneous measurements. Accordingly,~~ and therefore precipitation measurements during freezing temperatures are not available.

Two additional sources of climate data are available. Firstly, the Fagurhólsmýri weather station operated by the Icelandic Meteorological Office (IMO) approximately 12 km south of the study site has daily measurements of temperature dating back to 1949 and therefore provides long-term variations in temperature around the study region. The IMO have also recently produced a 2.5 km gridded dataset of total precipitation as part of the ICRA atmospheric reanalysis project (Nawri et al., 2017). ~~For this, they used the state-of-the-art HARMONIE-AROME mesoscale numerical weather prediction model (Bengtsson et al., 2017) forced by the latest ECMWF ERA-Interim reanalysis product from 1979-2016 to produce hourly precipitation data at a spatial resolution of 2.5 km over the whole of Iceland.~~ These data provide the best estimate of long-term precipitation at the study site and, given the limited availability of precipitation measurements at higher elevations around Öræfajökull, they also provide the best estimate of spatial variations in precipitation across the region.

2.2.2 Ice melt

An array of 17 ablation stakes installed by Flett (2016) in the main ablation zone of the glacier between 2012 and 2014 at elevations ranging from 142 to 462 m asl provide measurements of ice melt on the glacier tongue (Fig. 1). The BGS have also

undertaken annual high resolution (sub-meter) terrestrial LIDAR scans of the proglacial region including ice at the front of the glacier (see red dashed box in Fig. 1) which, given that ice flow is negligible here, provides an additional indication of ice melt.

Two digital elevation models of the ice also exist for the years 1988 and 2011 which indicate historical retreat of Virkisjökull. A 5 m 2011 DEM was constructed using high resolution airborne LIDAR scans of the ice surface (IMO, 2013) and is considered the most accurate measurement of the ice geometry and surrounding topography in the study region currently available. A 20 m 1988 DEM was derived from aerial photographs of the glacier (Landmælingar Íslands: www.lmi.is) using photogrammetric methods (Magnússon et al., 2016). Photogrammetry may suffer from errors due to image rectification and stereo-image mismatches (e.g. Barrand et al., 2009) and therefore the accuracy of this dataset is expected to be less certain, particularly over higher-elevation snow-covered terrain.

2.2.3 Snow coverage

~~Snow melt dynamics play an important role in the hydrological behaviour of mountainous catchments (Barnett et al., 2005; Jeelani et al., 20 and as such observation data relating to the accumulation and melt of snow are important for evaluating the performance of glacio-hydrological models.~~ No direct observations of snow accumulation or melt exist for the Virkisá river basin and so instead, satellite snow cover data (MOD10A1 product) from the Moderate Resolution Imaging Spectroradiometer (MODIS) (Riggs and Hall, 2015) were used. These data have been archived since 2000 and consist of daily 500 m gridded maps of snow cover extent with values ranging between 0 and 1 which relate to the proportion of the ground that is snow covered. While they do not provide a direct measurement of snow mass balance, they have shown to be a useful data source for evaluating the performance of GHMs ~~(Hanzer et al., 2016; Finger et al., 2015)~~(Finger et al., 2015; Hanzer et al., 2016). The quality of the data in high latitude regions such as Iceland are variable due to the need for good light and little or no cloud cover. As part of the MOD10A1 product, a basic estimate of the data quality is calculated as a means to avoid measurements affected by cloud cover and poor light conditions. For this study, only those data that achieved a QA score of 'good' or 'best' were used. This precluded the use of data collected between September to February presumably because of reduced daylight hours and increased cloud cover during these months.

2.2.4 River discharge

Hourly river discharge data collected since 2012 are available from an automatic stream gauge installed by the BGS 2 km downstream of the lake outlet on the Virkisá river (see ASG1 in Fig. 1). The gauge consists of two stilling wells with submerged pressure transducers which measure river stage and water temperature every 15 minutes. The stage data are subsequently converted into units of flow using a rating curve constructed from periodic river flow gaugings.

In conjunction with the river stage and water temperature measurements, a camera is mounted next to the river and takes photos of the channel three times a day. Given that the river is prone to freezing over the winter months, the photographic archive and temperature data were used to remove these periods from the river flow time-series. The river bed consists of large boulders (approximate diameter of 50 cm) which can become mobile during high flows causing shifts in the rating curve. For

this study, river discharge data for the years 2013 and 2014 were used as gauging for these years cover a wide range of flow magnitudes and rating shifts are limited and well constrained by observations.

2.3 Glacio-hydrological model

A distributed GHM which can incorporate different conceptual representations of melt and runoff-routing processes was used for all model experiments. The code was written in the object-oriented C++ programming language, making it computationally efficient and ideally suited for incorporating different model structures. The GHM consists of a 2D Cartesian grid of equally spaced model nodes. For this study, a spatial resolution of 50 m was selected as the best balance between simulation detail and model performance. Hourly observations of precipitation, temperature and incident solar radiation were used to simulate the accumulation of snowfall and the melt of snow, firn and ice across the model domain. The snow redistribution algorithm developed by Huss et al. (2008a) was used to account for snow drift and avalanches based on the curvature and slope of the surface. A soil infiltration and evapotranspiration model developed by Griffiths et al. (2006) solves the water balance for the non-glaciated regions of the study catchment. Excess soil moisture, rainfall and melt are then routed to the catchment outlet via a semi-distributed network of linear-reservoir cascades which represent the water storage and release characteristics of the major hydrological pathways in the watershed. The GHM also simulates the evolution of the glacier geometry under periods of sustained negative mass balance using the Δ-h parametrisation of glacier retreat which has shown to closely reproduce the evolution of Alpine glaciers with results comparable to more complex 3-D finite-element ice flow models ~~(Huss et al., 2010; Li et al., 2015; Van Tiel et al., 2017; Duethmann et al., 2016)~~(Huss et al., 2010). Details of this and the soil water balance component of the GHM can be found in Appendix A. The following text details the different melt and runoff-routing structures adopted for this study.

2.3.1 Snow and ice melt model structures

Melt of snow and ice is calculated at each model node separately. Snow melt can occur at any node where a snow pack has developed. Similarly, ice melt can only occur at ice-covered nodes where the snow pack has completely melted. The mass balance at a given node is the summation of snowfall minus snow and ice melt. The GHM uses the mass balance calculated at each node to determine the equilibrium line altitude (ELA) which is updated each simulation year. A rolling three-year average ELA was used to determine the dividing line between firn and ice on the glacier.

For this study, three different conceptual models of snow and ice melt with different levels of complexity were compared. All have been used extensively to simulate melt processes in glaciated regions around the world ~~(e.g Gao et al., 2017; Matthews and Hodgkins,~~ (e.g Matthews and Hodgkins, 2016; Ragettli et al., 2016; Gao et al., 2017; Nepal et al., 2017; Reveillet et al., 2017). The first melt model structure (TIM_1) employs a classic temperature index model approach (Braithwaite, 1995) whereby melt is assumed to increase linearly with temperature above a given critical threshold:

$$M_i = \begin{cases} a_i(T - T_i^*) & T > T_i^* \\ 0 & T \leq T_i^* \end{cases} \quad (1)$$

where a (m w.e. $^{o}C^{-1}$ h^{-1}) is the temperature factor calibration parameter that converts temperature into melt, T is the near-surface air temperature and T^* is the critical threshold above which melt occurs. To account for the different properties of snow, firn and ice that may bring about different values of a and T^*, these are defined separately so that $i = (snow, firn, ice)$.

The second melt model structure (TIM_2) was originally proposed by Hock (1999) which includes an additional incident solar radiation term to account for topographic effects such as slope, aspect and shading which ~~an~~ can bring about spatio-temporal variations in melt (Arnold et al., 2006; Pellicciotti et al., 2008). Their enhanced TIM has the form:

$$M_i = \begin{cases} (T - T_i^*)(a_i + b_i \cdot SW_{\downarrow}) & T > T_i^* \\ 0 & T \leq T_i^* \end{cases} \quad (2)$$

where b (m^3 w.e. W^{-1} $^{o}C^{-1}$ h^{-1}) is an additional radiation factor calibration parameter that converts the measured incident solar radiation, $SW_{\downarrow}$ (W m^2) into a unit melt. For this melt model structure the GHM accounts for shading using the DEM and position of the sun in the sky which is calculated for each hourly time-step using the SPA algorithm (Reda and Andreas, 2008). Additional perturbations in solar irradiance at the surface brought about by topographic effects such as slope and aspect are accounted for by calculating the incident angle of solar radiation to scale the measured incoming radiation.

Konya et al. (2004) noted that the form of Eq. (2) is not congruent with the full energy balance equation as temperature is used to multiply the shortwave radiation term which can lead to overestimation of melt during peak temperatures. Accordingly the melt model structure proposed by Pellicciotti et al. (2005) was also used for this study (TIM_3) which is an enhanced TIM in additive form that also incorporates an albedo parameter, α:

$$M_i = \begin{cases} a_i(T - T_i^*) + b_i \cdot SW_{\downarrow}(1 - \alpha_i) & T > T_i^* \\ 0 & T \leq T_i^* \end{cases} \quad (3)$$

where b has the units m^3 w.e. W^{-1} h^{-1}. Following Pellicciotti et al. (2005), this melt model structure also includes the dynamic snow albedo algorithm proposed by Brock et al. (2000) which accounts for the drop in snow albedo as it ages using a logarithmic function with the form:

$$\alpha_{snow} = p_1 - p_2 \cdot log_{10} \cdot T_a \quad (4)$$

where p_1 is the albedo of fresh snow (set to 0.9), p_2 is an empirical calibration parameter and T_a is the accumulated daily maximum temperature greater than $0^{o}C$ since snowfall.

For all melt model structures in the GHM, melt M is converted into a volumetric melt M_v at each node:

$$M_v = M \cdot A \quad (5)$$

where A is the model node area. Following Hopkinson et al. (2010) the area of each node is corrected for surface slope:

$$A = \frac{L^2}{\cos\beta} \quad (6)$$

where L is the model node length and β is the node surface slope.

2.3.2 Runoff-routing model structures

Runoff includes any rainfall falling on, and melting of the snow and ice as well as excess soil moisture from those areas free of ice and snow. The concept of linear reservoirs was employed to route this runoff to the catchment outlet. A linear reservoir receives a volumetric inflow and releases it at a rate proportional to its internal water storage following:

$$q = \frac{1}{k}s \tag{7}$$

where q is the outflow (m^3 h^{-1}), s is the storage (m^3) and k is mean residence time of the reservoir (h) which accounts for the diffusive effect of storage and release mechanisms within the catchment. Increasing the value of k increases the diffusion effect on the inflow hydrograph. Additional controls on the diffusion and lag effects can be obtained by arranging a cascade of multiple linear reservoirs in series (Ponce, 2014) so that the outflow from the previous reservoir is the inflow for the subsequent reservoir. With this setup, the continuity equation for the jth reservoir of n reservoirs in series, where $j = (1, 2...n)$ can be written as:

$$\frac{ds_j}{dt} = \begin{cases} i - q_j & j = 1 \\ q_{j-1} - q_j & j > 1 \end{cases} \tag{8}$$

The outflow hydrograph is then taken from q_n.

For temperate glaciers, the common practice is to subdivide the catchment into one or more hydrological response units (HRU) which are thought to have different water storage and release characteristics. For example, the firn, snow and bare ice have generally shown to respond over relatively long, intermediate and short time-scales respectively (Hock and Jansson, 2005) and therefore these may be characterised as separate HRUs, although as noted previously, simpler and more complex definitions of HRUs have been defined in the past. Subsequently, three runoff-routing model structures were proposed with different levels of complexity structured around these subdivisions (Fig. 2).

The first and simplest runoff-routing model structure (ROR_1) uses a single linear reservoir cascade (e.g. Boscarello et al., 2014) to route the inflow from all runoff sources simultaneously. This structure makes no distinction between the different runoff sources and flow pathways and assumes that all conform to the same storage-discharge relationship.

The second model structure (ROR_2), employs two linear reservoir cascades in parallel (e.g. Hannah and Gurnell, 2001). The first cascade represents the slow percolation of water through the snow and firn HRUs, while the second cascade represents faster flow of water through the bare ice and overland. This approach therefore makes some distinction between the different flow pathways and, by conditioning the parameters so that the snow and firn have a more diffuse response function, it introduces a degree of non-linearity in the discharge response to runoff.

The third runoff-routing model structure (ROR_3) has not been used previously. It employs separate linear reservoir cascades to route water from the firn, snow, ice and soil HRUs. Here the parameters are conditioned so that the firn is the most diffuse, slowly responding reservoir, followed by the snow and then the ice and soil zones are considered to be relatively flashy, quickly responding HRUs. This approach also includes some representation of linkages between these various units. Here it is hypothesised that water that flows through the firn, must then flow through the downstream bare ice HRU before it reaches the

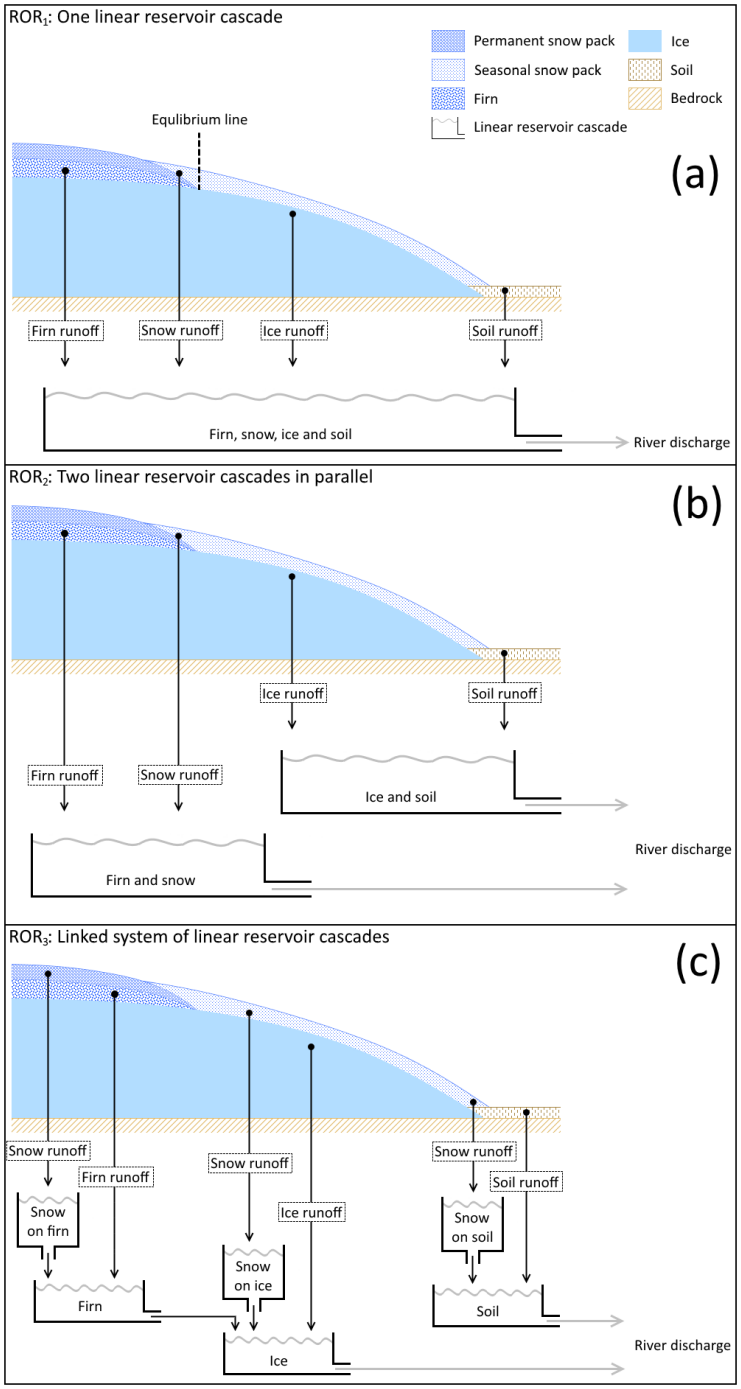

Figure 2. Three runoff-routing model structures which relate the linear reservoir cascade configurations to idealised cross-sections of a temperate glacier.

river. Similarly, water that percolates through the snow pack must also flow via the HRU that it overlies before it reaches the river. There are therefore six different flow pathways that runoff may take before reaching the river outlet (see Fig. 2c) and this represents the most complex, non-linear runoff-routing model structure.

2.4 Driving climate data

The GHM was configured to run from the initial ice geometry of 1988 through to 2015. It requires continuous measurements of hourly precipitation, near-surface air temperature and incident solar radiation to drive the various model components.

2.4.1 Precipitation

A new gridded precipitation time-series was constructed for the GHM that incorporates the measurements of rainfall from the weather stations in the Virkisá basin and the information on spatial and long-term variations in precipitation from the gridded ICRA reanalysis product. First, the weather station rainfall data were used to bias-correct the ICRA reanalysis product. Given that none of the weather stations are equipped with devices to measure snowfall, and that freezing temperatures can induce erroneous measurements in rainfall, only data with three consecutive preceding above-freezing days were used. This is a major issue for using AWS4 as the majority of days, particularly in the winter, are below freezing at this elevation. Accordingly, the AWS4 rainfall data were not used for the ~~bias correction~~ bias-correction procedure. Furthermore, because the AWS1 and AWS3 gauges overlap the same ICRA data pixel, and because the AWS1 time-series is the longest and most complete, it was decided to use the AWS1 data to bias-correct the overlapping ICRA data pixel. Here, the equidistant quantile mapping (EQM) approach ~~(Li et al., 2010; Srivastav et al., 2014; Sachindra et al., 2014)~~ (Li et al., 2010; Sachindra et al., 2014; Srivastav et al., 2014) was employed to bias-correct the ICRA precipitation time-series. EQM is an adaptation of the original quantile mapping method that accounts for non-stationarity in the moments of the biased time-series and helps to preserve changes in the cumulative distribution function of the precipitation data that may have occurred over time ~~(Switanek et al., 2017; Cannon et al., 2015)~~ (Cannon et al., 2015; Switanek et al., 2017). To evaluate the effectiveness of the bias-correction procedure, ~~the R^2 correlation score was calculated between the bias-corrected ICRA data and the measured~~ a number of statistics were calculated to compare the observed and ICRA precipitation data before and after bias-correction (Table 1). There were a total of 30,460 hourly measurements of precipitation available for above-freezing days at AWS1 ~~data. At~~ of which the majority were during the autumn months (September, October and November) and the least during the winter months (December, January and February). Overall, the procedure corrects for bias in the mean (Avg) and also improves the spread (SD), relative variability (CV) and skewness of the distribution of precipitation data at hourly, daily and 3-daily time-steps. At the seasonal scale, these improvements are notable for spring, summer and autumn. However, the bias-correction procedure typically has a slightly negative impact on the winter precipitation statistics, probably because of the limited above-freezing data available for these months. In particular, average hourly winter precipitation is underestimated by 0.11 mm (16%) while the positive bias in relative variability and skewness are amplified after bias-correction. Given that EQM preserves the rank correlation of the time-series, it has little effect on the R^2 correlation score, with a typical reduction of 0.01-0.02 after bias correction. At an hourly timescale, the ~~ICRA~~ bias-corrected data only captured 22% of the observed variance in the AWS1 rainfall record. However, when aver-

Table 1. Statistics calculated from the observed precipitation data at AWS1 and from the corresponding ICRA precipitation data before and after bias-correction. Statistics have been calculated at an hourly, daily and 3-daily time-step and include n (total number of above-freezing measurements available at AWS1), Avg (mean), SD (standard deviation), Cv (coefficient of variation), Skewness and R^2 (coefficient of determination).

Time-step	Statistic	Overall			Winter (DJF)			Spring (MAM)			Summer (JJA)			Autumn (SON)		
		Obs	Before	After	Obs	Before	After	Obs	Before	After	Obs	Before	After	Obs	Before	After
Hourly	n	30460			4344			6290			8832			10994		
	Avg (mm)	0.33	0.43	0.33	0.65	0.67	0.54	0.20	0.29	0.21	0.17	0.33	0.24	0.41	0.50	0.39
	SD (mm)	1.09	1.24	1.12	1.55	1.63	1.49	0.72	0.83	0.72	0.65	1.01	0.91	1.27	1.39	1.25
	Cv	3.28	2.85	3.37	2.39	2.42	2.78	3.59	2.82	3.46	3.79	3.02	3.74	3.11	2.79	3.23
	Skewness	5.81	5.40	6.01	3.84	4.38	4.83	5.91	4.82	5.43	7.45	6.41	7.51	5.43	4.93	5.35
	R^2		0.24	0.22		0.31	0.29		0.17	0.16		0.15	0.14		0.23	0.22
Daily	n	1264			181			260			368			455		
	Avg (mm)	7.95	10.4	7.93	15.6	16.2	12.9	4.84	7.02	5.00	4.13	7.99	5.86	9.78	11.9	9.31
	SD (mm)	16.2	19.5	17.1	23.2	27.2	24.4	10.0	12.2	10.2	8.80	13.1	11.3	18.9	22.6	20.0
	Cv	2.04	1.87	2.16	1.49	1.68	1.89	2.06	1.73	2.04	2.13	1.64	1.92	1.93	1.89	2.15
	Skewness	4.37	4.67	5.07	2.78	4.50	4.66	3.71	3.38	3.92	4.39	2.64	2.92	4.19	4.02	4.36
	R^2		0.51	0.49		0.64	0.62		0.44	0.43		0.49	0.47		0.45	0.43
3-daily	n	385			47			75			123			140		
	Avg (mm)	23.5	31.0	23.6	48.6	50.1	40.0	13.7	19.8	14.0	12.4	23.8	17.4	30.0	36.8	28.7
	SD (mm)	34.1	41.6	36.4	47.7	67.4	60.9	18.4	23.9	19.5	18.2	27.1	22.8	39.7	45.4	39.7
	Cv	1.45	1.34	1.54	0.98	1.35	1.52	1.35	1.21	1.39	1.46	1.14	1.31	1.32	1.23	1.38
	Skewness	2.94	3.44	3.78	1.86	3.36	3.45	1.96	1.92	2.22	2.82	1.84	1.98	2.58	2.28	2.53
	R^2		0.73	0.72		0.79	0.77		0.60	0.59		0.56	0.54		0.68	0.66

aged to a daily timescale the R^2 score increased to 0.49, and for a three-daily ~~average~~ timescale the R^2 increased to 0.72. The limited ~~accuracy~~ correlation of the ICRA precipitation data at an hourly timescale could hinder the acceptability of the GHM across some of the signatures (e.g. the river discharge signatures related to the timing of flows). However, the AWS1 rainfall record is complete for the years 2013 and 2014 where the GHM is compared against observed river discharge signatures. As such, poor replication of the timing of hourly rainfall events should ~~not influence~~ have minimal influence on the GHM's ability to capture the river discharge signatures. Rather, the role of the bias-corrected ICRA precipitation data was primarily to drive the glacier-mass balance component of the GHM prior to 2009 for which ~~reliable precipitation data on a~~ a reliable three-daily ~~timescale were~~ temporal correlation with observations was deemed adequate.

2.4.2 Near-surface air temperature

The longest record of hourly temperature measurements in the Virkisá river basin are from AWS1 which starts in 2009. To generate a continuous time-series of temperature back to 1988, daily measurements of temperature available from the nearby Fagurhólsmýri weather station were used. A comparison of daily average temperatures showed there to be a good linear relationship between the two stations with an R^2 of 0.92. As such, this linear model was used to bias-correct the daily weather station data so that it could be combined with the AWS1 time-series. To downscale the data to an hourly resolution,

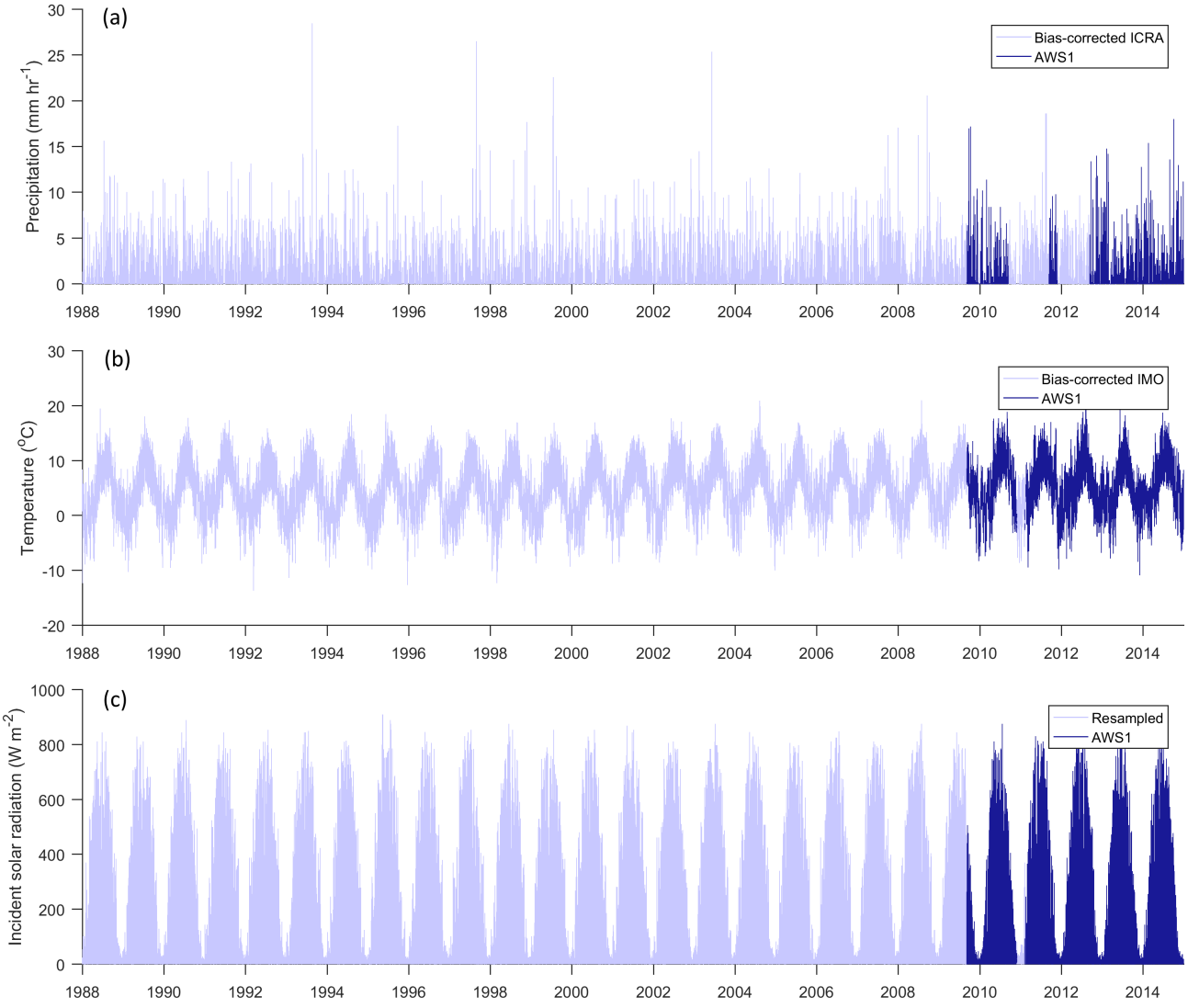

Figure 3. Continuous hourly time-series of precipitation (a), temperature (b) and incident solar radiation (c) between 1988 and 2015 at AWS1.

24-hour temperature anomalies were randomly sampled from the AWS1 record, thereby ensuring the complete time-series had a consistent sub-daily variability. Of course, diurnal cycles in temperature are dependent on the time of year, whereby increased incident solar radiation in the summer enhances sub-daily temperature variability. Therefore, the sampling strategy was employed on a month-by-month basis. The complete hourly time-series of temperature at AWS1 is shown in Fig. 3b.

As in many glaciated catchments topography ~~, to a large extent,~~ controls spatial temperature variations to a large extent. The importance of characterising temperature lapse rates for glacio-hydrological modelling is well known because it has a strong control on spatial patterns of melt simulations ~~(Gardner and Sharp, 2009; Heynen et al., 2013; MacDougall et al., 2011)~~ (Gardner and Sharp, 2009; MacDougall et al., 2011; Heynen et al., 2013). In fact while many studies employ a fixed temperature lapse rate, in reality seasonal variations in surface characteristics (e.g. albedo and roughness) and atmospheric conditions can bring about strong seasonal and diurnal variations in lapse rates which control melt processes (Gardner et al., 2009; Minder et al., 2010; Immerzeel et al., 2014). Furthermore, local atmospheric phenomena associated with mid-latitude glaciers such as katabatic winds which bring cool dense air over the ice surface can serve to ~~shallow~~ reduce the temperature gradient (Petersen and Pellicciotti, 2011; Ragettli et al., 2014). Having analysed near-surface air temperature variations both on and away from the Virkisjökull glacier, it was deemed most appropriate to extrapolate temperature across the study catchment using a seasonally variable hourly lapse-rate in conjunction with an on-ice temperature correction function based on the work of Shea and Moore (2010) (see Appendix B).

2.4.3 Incident solar radiation

The only source of incident solar radiation is the continuous hourly time-series from AWS1. To construct a continuous time-series back to 1988, a resampling strategy was employed to generate a complete time-series that was statistically consistent with the data at AWS1. It was found that during the summer months, the daily range in incident solar radiation and temperature are strongly correlated. Therefore, when generating a continuous time-series of hourly incident solar radiation from 1988, it was important to maintain this dependence between intra-day solar radiation and temperature variability. To do this, a coordinated (in time) sampling strategy identical to that used for the near surface air temperature data was employed. More specifically, for each random 24-hour temperature anomaly sample from the AWS1 record used to build part of the temperature time-series, the corresponding 24-hour solar cycle data were extracted and used to build the same part of the incident solar radiation time-series. Figure 3c shows the complete time-series of incident solar radiation used to drive the model.

2.5 Signatures and limits of acceptability

Observations of ice melt, snow coverage and river discharge were used to derive 33 unique signatures with LOA to characterise the glacio-hydrological behaviour of the Virkisá river basin over different spatio-temporal scales and evaluate the acceptability of the different model structures (Table 2). For convenience, the signatures have also been subdivided into 11 attributes which encapsulate the main aspects of model behaviour that were assessed.

Table 2. Summary of signatures used to evaluate model acceptability. Units with asterisk (*) are per section of flow duration curve.

Group	Attribute	Attribute ID	Signature	Limits of acceptability
Ice melt	Seasonal ice melt on tongue	Seas melt	2013 Summer ice melt	5.22 – 6.44 m we
			2012-2013 Winter ice melt	0.64 – 0.78 m we
	Long term glacier volume change	Melt vol	Change in ice volume (1988-2011)	-0.36 – -0.28 km^3
Snow coverage	Snow coverage in lower catchment	Low snow	Mean snow coverage in spring	0.32 – 0.45
			Mean snow coverage in early summer	0.02 – 0.08
			Mean snow coverage in late summer	0.00 – 0.03
	Snow coverage in mid catchment	Mid snow	Mean snow coverage in spring	0.70 – 0.80
			Mean snow coverage in early summer	0.17 – 0.27
			Mean snow coverage in late summer	0.00 – 0.04
	Snow coverage in upper catchment	Upp snow	Mean snow coverage in spring	0.81 – 0.90
			Mean snow coverage in early summer	0.51 – 0.64
			Mean snow coverage in late summer	0.02 – 0.09
River discharge	Mean monthly river flow	Mnthly flow	Mean January river flow	1.16 – 1.86 $m^3\ s^{-1}$
			Mean February river flow	1.69 – 2.92 $m^3\ s^{-1}$
			Mean March river flow	0.85 – 1.58 $m^3\ s^{-1}$
			Mean April river flow	0.73 – 1.48 $m^3\ s^{-1}$
			Mean May river flow	1.50 – 2.16 $m^3\ s^{-1}$
			Mean June river flow	4.12 – 6.23 $m^3\ s^{-1}$
			Mean July river flow	6.33 – 10.30 $m^3\ s^{-1}$
			Mean August river flow	5.72 – 9.15 $m^3\ s^{-1}$
			Mean September river flow	4.55 – 7.38 $m^3\ s^{-1}$
			Mean October river flow	3.88 – 7.02 $m^3\ s^{-1}$
			Mean November river flow	3.90 – 7.40 $m^3\ s^{-1}$
	Quick release high flows	High flows	Volume under highest flow section of FDC	59.4 – 116.0 $m^3\ s^{-1}$ *
			Slope of highest flow section of FDC	2.67 – 9.88 $m^3\ s^{-1}$ *
			Volume under high flow section of FDC	70.6 – 111.0 $m^3\ s^{-1}$ *
			Slope of high flow section of FDC	0.38 – 0.79 $m^3\ s^{-1}$ *
	Slow release low flows	Low flows	Volume under low flow section of FDC	20.9 – 46.1 $m^3\ s^{-1}$ *
			Slope of low flow section of FDC	0.03 – 0.05 $m^3\ s^{-1}$ *
	Flow variability	Flow var	Coefficient of variation	0.95 – 1.83
	Melt runoff timing	Melt timng	Peak summer flow hour	17:00 – 18:00
	Flashiness	Flow flash	Integral scale	25 – 44 h
			Rising limb density	0.13 – 0.20

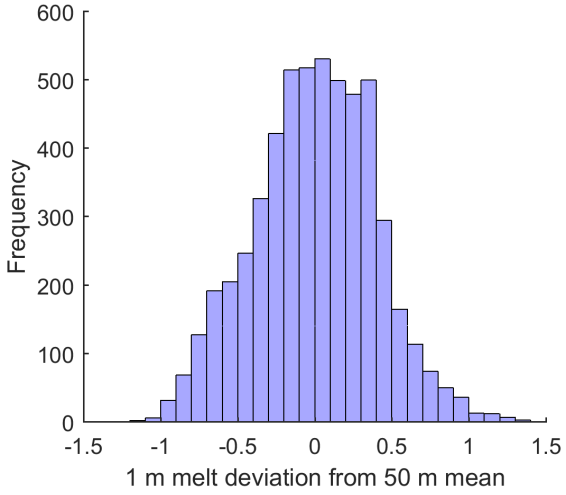

Figure 4. Histogram of deviation of 1 m melt from 50 m mean derived from terrestrial LIDAR scans of static ice front between 2012-2014.

2.5.1 Ice melt

The average winter (November 2012 - April 2013) and summer (May 2013 - September 2013) melt across the ablation stake network were used to characterise the short-term, seasonal ice melt on the glacier tongue. Of course, point measurements of melt are not directly comparable to simulated melt at the GHM nodes as these simulations represent the average melt over the node area. Therefore, the GHM can only be expected to get as close to the stake measurements as the actual spread in melt over the equivalent model node area. To calculate this spread, the high resolution terrestrial LIDAR scans taken during the ablation stake campaign were used. The scans were used to estimate the spread of melt deviations from the mean melt across 50 m square regions (Fig. 4). The 95% confidence bounds ($\pm$ 0.78 m yr^{-1}) were then used to define the LOA around the winter and summer melt signatures where it was assumed that the spread should be proportional to the total melt. This assumption leads to much narrower LOA around the winter melt signature than the summer melt signature.

A signature to characterise the long-term change in glacier volume was also quantified by differencing two 3D models of the ice from 1988 and 2011. These models were constructed using the two ice surface DEMs in combination with a bedrock model of the Öræfajökull region (Magnússon et al., 2012). Given the potential errors in the 1988 DEM, this dataset was assumed to be the main source of uncertainty in the calculation of the ice volume change signature. A comparison to the more accurate 2011 DEM shows that the 1988 DEM captures the gridded elevation data across the non-glaciated portion of the study area with reasonable accuracy (Fig. 5a). The residuals are approximately normally distributed with a mean error of zero (Fig. 5b) and they show to be largest for those parts of the catchment that are steeply sloped (scatter in Fig. 5c). To account for these errors in the calculation of the ice volume change signature, 1000 unique DEMs of the 1988 ice surface were generated by randomly perturbing each pixel of the original dataset with perturbations drawn from a normal distribution with mean zero. Given that the spread of the residuals increases for those areas of the catchment that are steepest, the shape parameter of the

error distribution (standard deviation) was varied according to the slope of each pixel of the 1988 DEM (see dark blue line in Fig. 5c). From these, 1000 equally probable estimates of ice volume change were calculated and the 95% confidence interval was used to define the LOA. The total change in ice volume over 23 years from 1988 was estimated to be between -0.36 and -0.28 km^3.

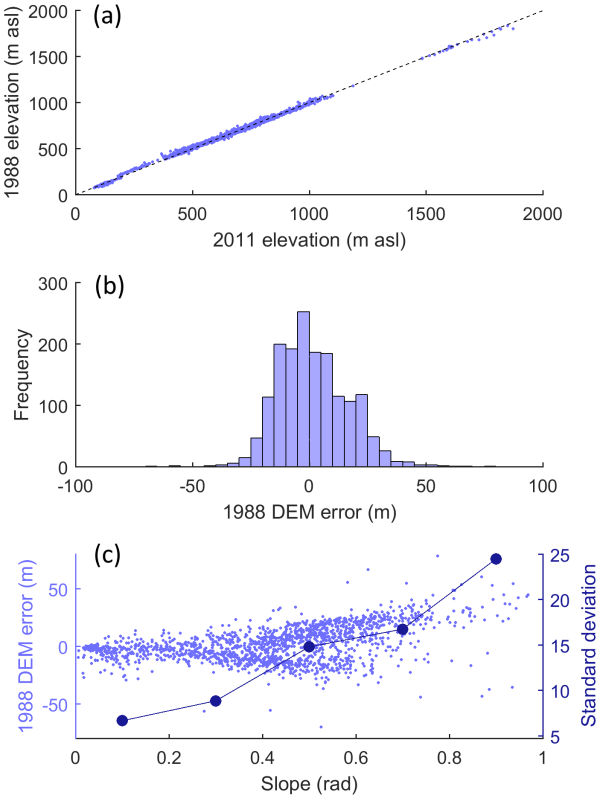

Figure 5. Error model for estimating uncertainty in glacier volume change between 1988 to 2011 including: 1988 vs 2011 off-ice DEM elevations (a), distribution of 1988 DEM errors calculated as difference between 1988 and 2011 off-ice elevations (b) and estimation of change in standard deviation of errors with DEM slope (c).

2.5.2 Snow coverage

Having removed the MODIS data that did not pass the QA test including all of the data between September and February, less than 5% of the remaining data were usable, and therefore, it was decided that these data should be combined to derive three seasonal average snow coverage maps. From these maps, three snow coverage curves were constructed that define the mean catchment snow coverage over an elevation range for three different times of the year: spring (March and April), early summer (May and June) and late summer (July and August) (Fig. 6). The curves ~~,~~ provide information on both the spatial and temporal distribution of snowfall in the study catchment. They were constructed by distributing the seasonal average snow distribution

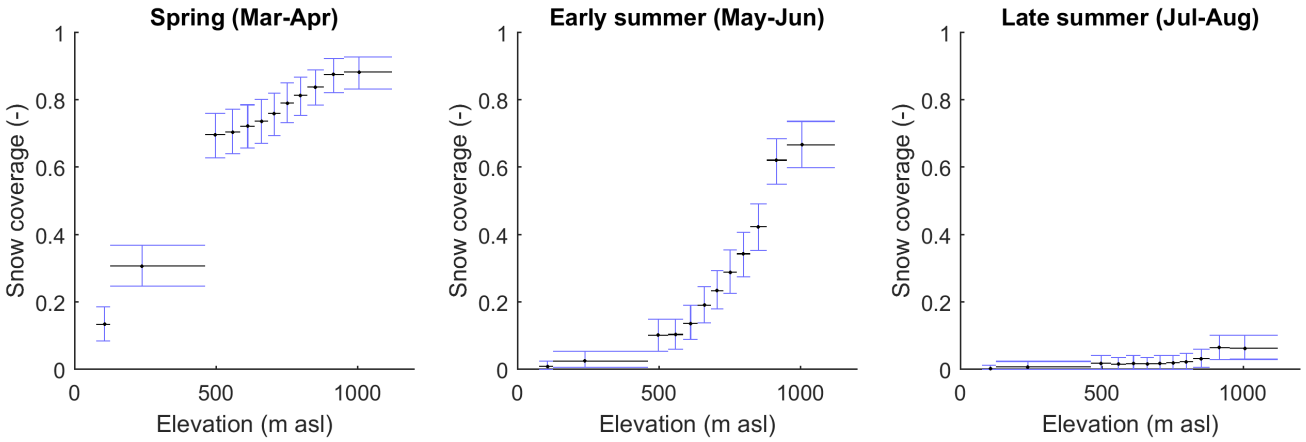

Figure 6. Snow coverage curves defined from the MOD10A1 snow cover product from 2000 - 2015 with 95% confidence bounds.

maps across the 50 m model grid DEM. For example, for a MODIS pixel value of 0.5, 50 of the corresponding DEM pixels were assumed to be snow covered. The MOD10A1 product cannot distinguish between snow and ice-covered regions, so only data that covered ice-free parts of the catchment were used. This limited the analysis up to a maximum elevation of just under 1200 m asl. While this does not cover the full elevation range of the catchment, Fig. 6 shows that the three curves capture a large amount of variability in seasonal snow cover. From the three snow coverage curves, the mean snow coverage from the lower, middle, and upper terciles of the curves were used as signatures of snow coverage.

There exists no definitive quantification of errors in the MOD10A1 product that can be used to estimate LOA for these signatures. Previous validation of the MODIS data using satellite imagery has shown the data to be relatively robust (Salomonson and Appel, 2004). Accordingly, it was assumed that as with the ablation stake data, the primary source of uncertainty stems from scale differences between the data and the model simulations. More specifically, because the MODIS data have a coarser resolution (500 m) than the DEM over which the MODIS data were distributed (50 m), a MODIS pixel value of 0.5 only indicates that 50 of the corresponding 100 DEM pixels are snow covered. The construction of a snow distribution curve, therefore necessitates some assumptions about where the snow actually lies which will influence the shape of the snow distribution curve. Accordingly, the LOA were quantified to account for this uncertainty. Here, for each of the seasons, a mean MODIS snow cover map over the study region was derived. Then, for each 500 m pixel, snow was randomly distributed across the corresponding DEM pixels 1000 times. From these, an equal number of snow distribution curves and corresponding snow distribution signatures could be derived, each assumed to be equally probable. The 95% confidence bounds from this distribution of snow cover signatures were used to define the LOA which are indicated by blue error bars in Fig. 6.

2.5.3 River discharge

The hourly river discharge data for the years 2013 and 2014 measured at ASG1 (Fig. 7a) were used to define 21 different river discharge signatures that cover a range of temporal scales and flow magnitudes. The majority of these signatures were based on previous studies ~~(Coxon et al., 2014; Yilmaz et al., 2008; Westerberg et al., 2016; Shafii and Tolson, 2015; Hrachowitz et al., 2014; Schaef~~ ~~and are summarised as follows~~(e.g Yadav et al., 2007; Yilmaz et al., 2008; Shafii and Tolson, 2015; Schaefli, 2016).

Mean monthly river flows were calculated to characterise the seasonal river flow regime. Signatures were also derived from sections of the flow duration curve to characterise quick-release high flows and slow-release low flows. These include signatures that quantify the volume under the section (flow magnitude) and the slope of section (flow variability) for the low flow section (99-66% flow exceedance), high flow section (15-5% flow exceedance) and highest flow section (5-0.5% flow exceedance). An overall estimate of flow variability, the coefficient of variation, was also calculated. Related to this, two further signatures, the rising limb density and integral scale, provide a measure of flashiness. The rising limb density is the ratio of number of flow peaks to the total time to peak where a higher number is more flashy. The integral scale measures the lag time at which the autocorrelation function of the flow time-series falls below $\frac{1}{e}$ (diurnal cycles in river flow were removed prior to this using a moving average filter). A higher integral scale therefore indicates a more slowly responding hydrological system. Finally, the peak summer flow hour of the observed discharge time-series was calculated to characterise the intra-day river discharge response to melt.

Estimates of river discharge are inherently uncertain (Pappenberger et al., 2006). McMillan and Westerberg (2015) provide a useful definition of two important sources of uncertainty which they distinguish as either aleatory (random) or epistemic (of an unknown character). The first stem from random measurement errors such as those from the instrument used for periodic river gaugings. These cause gauging points to vary around the 'true rating curve', typically according to some formal statistical definition. Epistemic uncertainty stems from the assumptions hydrologists have to make when constructing rating curves such as assuming the river bed profile and horizontal flow velocity distribution is relatively stable over time. These errors make fitting a single rating curve to all of the gauging data invalid. Accordingly, McMillan and Westerberg (2015) propose a method to define rating curve uncertainty which accounts for both sources of error and has been used to estimate uncertainty in river discharge signatures (Westerberg et al., 2016). The random error component was defined from analysis of 27 flow gauging stations in the UK with stable ratings and without obvious epistemic errors (Coxon et al., 2015). They conclude that this source of error is best approximated by a logistical distribution model. To account for epistemic error, they reject the assumption that the rating curve is fixed in time and instead they fit an ensemble of rating curves to all of the gauging data. Each curve is weighted by a 'Voting Point' likelihood function which scores it based on how many points of the periodic gaugings it is able to intersect (and at what location in the logistical distribution of each measurement).

In this study, the methodology proposed by McMillan and Westerberg (2015) was used to estimate rating curve uncertainty. Markov chain Monte Carlo sampling was used to define 667 unique rating curves which together define the rating curve uncertainty (Fig. 7b). From these an equivalent distribution of each river discharge signature was derived from the ensemble of flow time-series (Fig. 7c), from which the 95% confidence bounds were used as the LOA. Because the Voting Point method

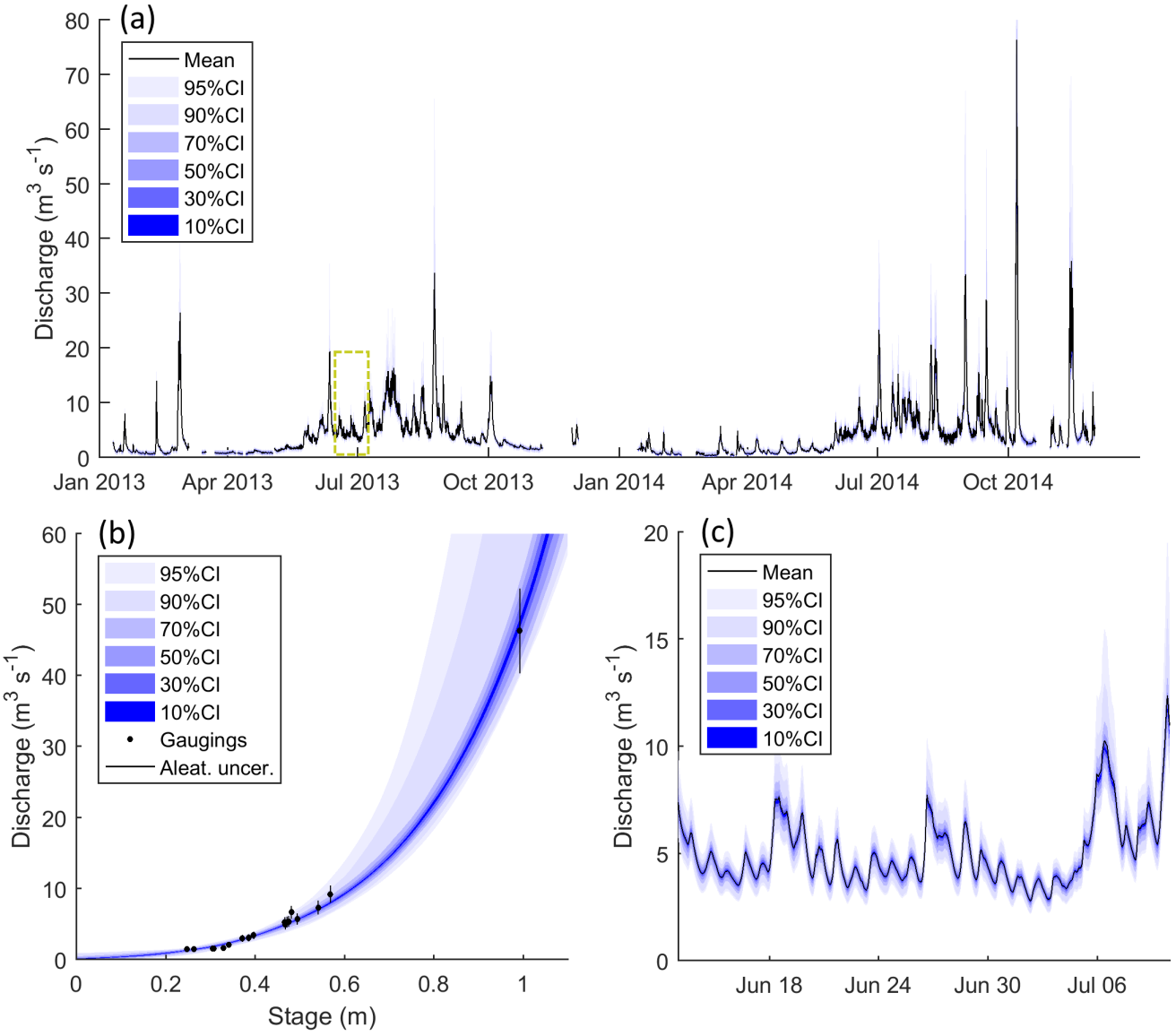

Figure 7. River flow time-series from ASG1 with quantified confidence intervals (a), rating curve uncertainty used to quantify confidence intervals (b) and zoomed section of river flow time-series (see yellow dash box in top plot) with confidence intervals (c).

only accounts for uncertainty in the flow magnitude and not the timing, it was not suitable to apply this approach to the three signatures that characterise melt runoff timing and flashiness. For these signatures, Schaefli (2016) proposed that the LOA should be derived by subsampling different periods of the flow time-series. For this study a month-by-month sub-sampling strategy was employed to do this.

2.6 Model calibration procedure

The GHM was configured to run from 1988 to 2015 so that simulations could be compared against all observation signatures. The initial ice surface was set to the 1988 DEM of the ice while the bedrock and land surface topography were taken from the Öræfajökull bedrock map (Magnússon et al., 2012). Initial snow coverage, soil moisture, linear reservoir storages and ELA were determined by running the model for three consecutive years prior to the simulation period using climate data from 1985 to 1988.

In total there were nine possible structural configurations of the GHM including all possible combinations of the three melt and runoff-routing model structures. For each of the nine configurations, the melt and runoff-routing model parameters were calibrated to achieve the closest fit to the observed signatures. To do this, first a set of preliminary runs were undertaken to assess the sensitivity of the simulations to the parameters. Here, it was found that the simulations were insensitive to the firn melt parameters across the range of 33 signatures. Accordingly, these were set to the same values as for snow. Similarly, none of the signatures were sensitive to the threshold above which melt occurs, T^*, and accordingly, this was set to 0 ^{o}C throughout the model experiments. Finally, it was also decided to fix the albedo parameter for ice in TIM_3 to 0.3. This was because this parameter directly interacts with the b parameter and therefore provides no extra control over model behaviour.

The remainder of parameters were kept for calibration (see Table C1). For each GHM configuration, 5000 Monte Carlo simulations with random parameter sets sampled from pre-defined uniform distributions were undertaken. The prior parameter distributions were defined from a review of previous modelling studies and later refined during the preliminary runs noted above. The quasi-random Sobol sampling strategy (Brately and Fox, 1988) was employed to sample the parameter space as efficiently as possible. The simulated signatures from each model run (parameter set) were then evaluated against the observed signatures using a continuous acceptability score:

$$s_j = \begin{cases} 0 & low_j \leq sim_j \leq upp_j \\ \frac{sim_j - upp_j}{upp_j - obs_j} & sim_j > upp_j \\ \frac{sim_j - low_j}{obs_j - low_j} & sim_j < low_j \end{cases} \quad (9)$$

where obs_j and sim_j are the observed and simulated values for signature j and upp_j and low_j are the upper and lower LOA. Here, a score of zero indicates that the model captures the observed signature within the LOA. An absolute score greater than 0 is outside of the LOA and therefore unacceptable. The sign of the score indicates the direction of bias while its magnitude indicates the model's performance relative to the LOA. A score of -3 would indicate that the model underestimates the signature by three times the observation uncertainty.

Given that there are 33 different signatures to calibrate to simultaneously, it was important to define a weighting scheme to achieve the best overall performance across the range of signatures. It was decided that, for a given GHM configuration, the 5000 runs should be ranked by a weighted average score where each group, each attribute within each group and each signature within each attribute were given equal weighting so that the scores were not biased to a particular group or attribute. The top 1% of model runs that achieved the smallest weighted average acceptability scores were then taken as the calibrated models for each GHM configuration and the average acceptability scores of these are reported. A bootstrapping with replacement re-sampling scheme was also used to assign 95% confidence intervals around all reported acceptability scores. While not a formal test of statistical significance, these were used to avoid reporting differences between the GHM configurations where issues such as under-sampling of the parameter space would make such conclusions unjustified. Where confidence intervals do not overlap, differences are hereafter referred to as substantial. The different GHM configurations were also compared when calibrated to individual groups of signatures (ice melt, snow coverage and river discharge). In this case the same weighting procedure was applied to a single group only.

3 Results

3.1 Signature discrimination power

As a first step towards evaluating the LOA framework, the discrimination power of the signatures was investigated to determine their relative usefulness for discriminating between acceptable and unacceptable model structures and parameterisations when used individually. A total of 45,000 calibration runs, each with unique model structures and parameterisations (hereafter referred to as model compositions) were undertaken in this study. The signatures with the highest discrimination power were defined as those that best constrain the range of acceptable model compositions. Here, the total number of acceptable model compositions were calculated for each signature as an indicator of discrimination power (bars in Fig. 8a). The results indicate that the ice melt signatures are the best discriminators, where each accepted less than 5000 model compositions. Of these, the winter melt signature from the ablation stake measurements is the best discriminator while the summer melt signature shows the least discrimination power. The snow coverage signatures generally show to be inferior discriminators when compared to the ice melt signatures. The late summer snow coverage signature for the lower catchment shows to be the poorest discriminator, presumably because there is negligible snow cover here at this time of the year; an observation that almost all of the model compositions have no difficulty in replicating. In contrast, no model compositions are deemed acceptable for the signatures of the spring and early summer snow coverage in the upper catchment. ~~This could indicate a deficiency in the melt model structures tested in this study and is considered in the remainder of the analysis.~~

The discrimination power of the river discharge signatures shows to be highly variable, but there are several discernible patterns. Firstly, the mean monthly flow signatures between January and June, when river discharge is low, show to be better discriminators than the higher-flow signatures from July to October. The mean monthly January and May flows stand out as being particularly powerful at discriminating between acceptable and unacceptable model compositions suggesting that these are likely to be important focal points for characterising model deficiencies. Those signatures related to the variability of flows

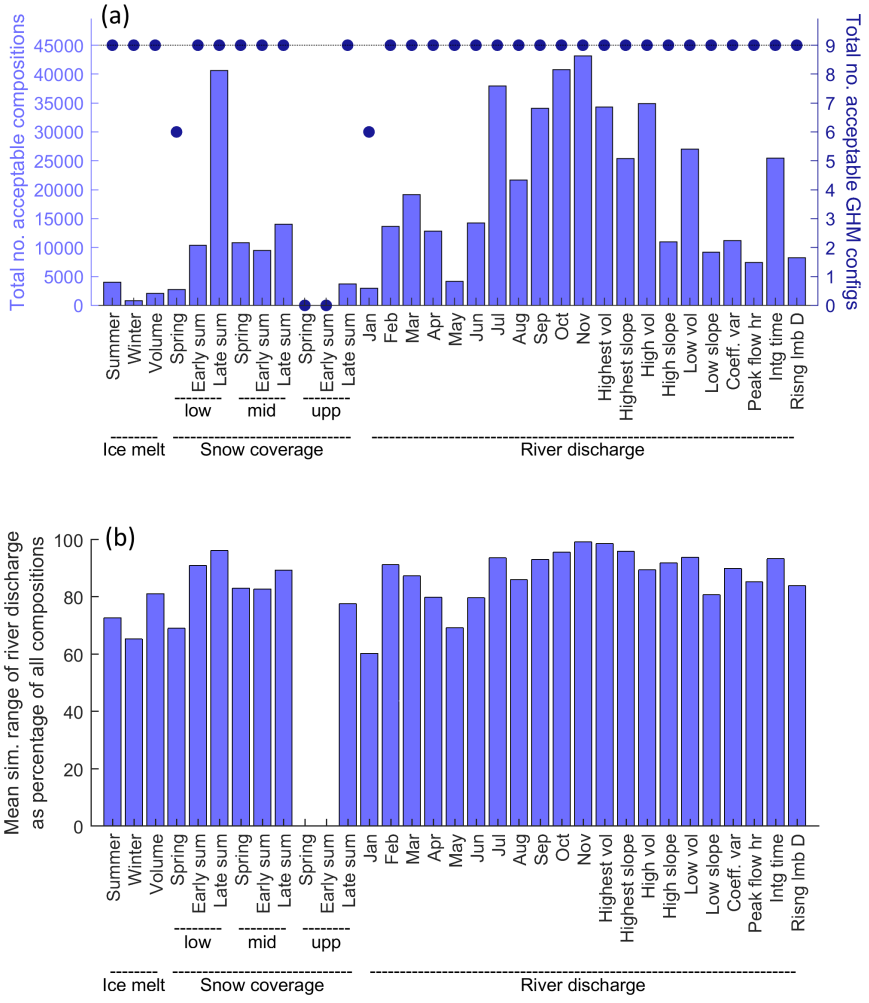

Figure 8. Total number of acceptable model compositions (bars) and configurations (dots) for each signature (a) and mean simulated range in river discharge from the population of acceptable models as a percentage of the simulated range using all of the 45000 model compositions (b).

such as the coefficient of variation and the flow duration curve slope signatures, as well as peak flow hour (timing) and rising limb density (flashiness) also show to be relatively good discriminators.

To determine the structural discrimination power of each signature, the total number of GHM configurations that returned at least one acceptable simulation has also been calculated for each signature (scatter in Fig. 8a). They show that when used individually, most of the discrimination power stems from constraining the parameter space rather than constraining the structural space. Only the lower-catchment spring snow coverage and mean January river flow signatures discriminate between structures where only six of the nine GHM configurations returned acceptable simulations. In both cases it was the GHM configurations that employed the TIM_3 melt model structure that could not capture these signatures within their LOA.

To indicate of how each signature helps to reduce river flow prediction uncertainty, a second measure of discrimination power has also been calculated (Fig. 8b). Here, the mean simulated range in river discharge from the population of acceptable models has been calculated as a percentage of the simulated range using all of the 45,000 model compositions for each signature. These results show that when used individually, all of the signatures help to constrain the river flow prediction uncertainty, although the effectiveness of each is variable. The mean January and May river flow signatures again exhibit good discrimination power, reducing the mean river discharge uncertainty to 60-70% of that from the full population of model compositions. Similarly, the winter ice melt and spring snow coverage in the lower catchment remain as two of the best discriminators. However, some signatures such as the long-term volumetric change in the glacier, which showed to be a good discriminator of model acceptability, are not as effective at reducing river discharge prediction uncertainty. ~~Overall, these results highlight the contrasting discriminatory power of the different signatures employed in this study.~~

3.2 Acceptability of melt model structures

While all signatures clearly demonstrate discrimination power when used individually, it remains to be seen how effective the LOA framework is for discriminating between and diagnosing deficiencies in different model structures when using multiple evaluation criteria. Here, the acceptability scores obtained after calibrating the GHM to the different groups of signatures (ice melt, snow coverage and river discharge) using the three different melt model structures have been calculated (Fig. 9). The light grey boxes indicate those signatures that have been captured within the LOA and the dark grey boxes and their corresponding acceptability scores indicate those signatures for which the structures were not able to capture within the LOA. So that the river discharge acceptability scores can be compared fairly, they have all been obtained using the ROR_1 runoff-routing structure.

When calibrated against the ice melt signatures, the GHM is not able to capture them within their LOA, regardless of the melt model structure used. The different GHM configurations show a tendency to overestimate the measured summer and winter melt from the ablation stake data, yet underestimate the long-term change in total ice volume (note underestimation here refers to the simulated loss in ice volume). This highlights a deficiency in the melt model structures as they are unable to reconcile the three melt signatures simultaneously within the observation uncertainty. The winter melt is by far the most unacceptable simulation, particularly when using the TIM_1 structure where it is overestimated by more than 30 times the observation uncertainty.

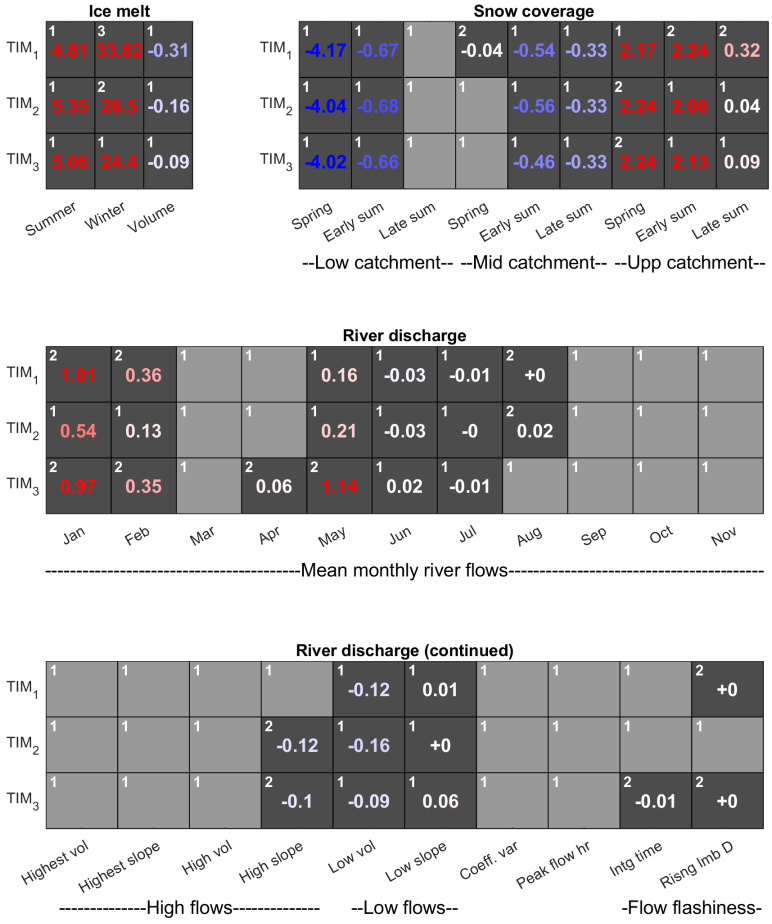

Figure 9. Acceptability scores obtained after calibrating the GHM using the three melt model structures in combination with the ROR_1 runoff-routing model structure. The three GHM configurations were calibrated against ice melt, snow coverage and river discharge signatures separately. Light grey boxes indicate acceptable simulations ($s = 0$) and numbered, dark-grey boxes indicate unacceptable simulations coloured blue and red to indicate negative and positive bias respectively. Note, all acceptability scores are rounded to two decimal places. Those non-zero scores that round to zero are accompanied by +/- to indicate sign of score. White numbers in top left of each box indicate relative ranking where acceptability scores are substantially different between the GHM configurations.

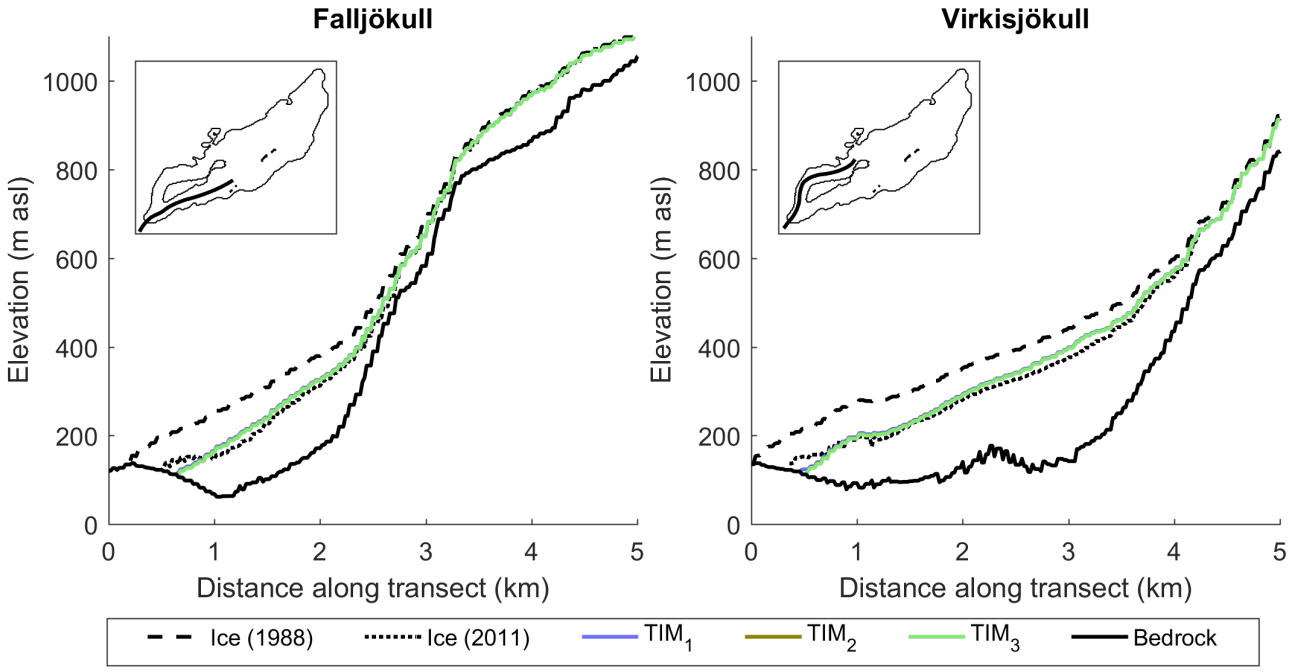

Figure 10. Observed and simulated ice thickness change as measured along transects of the Falljökull and Virkisjökull glacier arms. Insets show transect location.

Each of the GHM configurations using the three melt model structures have been ranked from 1 to 3 in the top left corner of each box where the acceptability scores are substantially different (Fig. 9). While there are clearly differences in the acceptability scores for the summer melt and ice volume signatures, these are not substantially different and therefore it is not possible to say that one structure is more acceptable than another. Indeed, a comparison of the simulated ice thickness change along the Falljökull and Virkisjökull arms of the glacier reveal that all three configurations of the GHM produce almost identical simulations which broadly capture the observed ice thickness change between 1988 to 2011 (Fig. 10).

For the winter melt signature, there is a substantial difference in acceptability when using the three melt model structures. Here, the GHM configuration using the TIM_3 structure is the most acceptable while that using the TIM_1 structure is least acceptable, indicating that while all configurations produce simulations outside of the LOA, there is an improvement in ice melt simulations when implementing the most sophisticated TIM_3 melt model structure.

For the snow coverage signatures, all three of the GHM configurations capture the late summer snow coverage in the lower portion of the catchment within the LOA. When using the TIM_2 and TIM_3 structures the mid-catchment spring snow coverage is also captured. The remaining snow coverage signatures are not captured within the LOA where all configurations show a tendency to underestimate snow coverage in the lower and mid parts of the catchment and overestimate snow coverage in the upper part of the catchment. To investigate why this is, Fig. 11 (left) shows the simulated early summer mid-catchment and upper-catchment snow coverage signatures for the 5000 calibration parameter sets (blue dots) used with the TIM_1-ROR_1 GHM

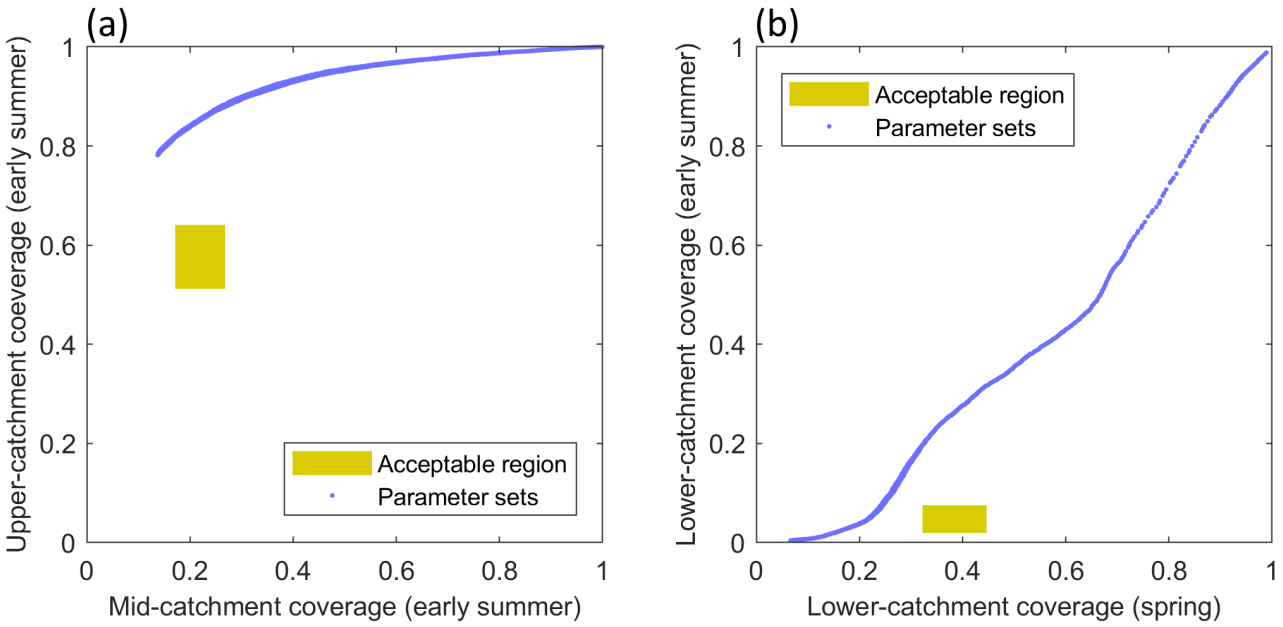

Figure 11. Simulated snow coverage signatures from the 5000 calibration runs (blue dots) for the TIM_1-ROR_1 GHM configuration including: early summer mid-catchment and upper-catchment snow coverage signatures (left), and lower-catchment spring and early-summer snow coverage signatures (right).

configuration. Here it can be seen that regardless of the choice of melt model parameters, this structure is not able to capture both of these signatures within their LOA simultaneously (indicated by yellow area). A similar inconsistency exists when comparing snow coverage over different seasons where the GHM is not able to capture the lower catchment snow coverage in the early summer and spring simultaneously (Fig. 11, right). Indeed, this inconsistency extends across all melt model structures.

A comparison of simulated snow distribution curves from the calibrated models (Fig. 12) reveals that all return similar simulations. The simulation using TIM_1 deviates slightly from the curve produced by the GHM when using the TIM_2 and TIM_3 structures, but overall the choice of melt model structure has a limited influence on the simulated seasonal snow coverage.

The acceptability scores for the river discharge signatures in Fig. 9 show that regardless of the choice of melt model structure, when used in conjunction with the ROR_1 runoff-routing model structure, all are able to capture a range of the river discharge signatures. The simplest GHM configurations using the TIM_1 and TIM_2 model structures capture 12 river discharge signatures simultaneously within the LOA while the inclusion of the dynamic snow albedo term and re-arrangement of the melt equation in the TIM_3 melt model actually inhibits the GHM performance where only 10 of the 21 river discharge signatures are captured within the LOA.

The mean monthly flow signatures for January, February and May show some of the highest absolute acceptability scores indicating the models are least efficient at capturing these. For winter flows in January and February, the simulation using the

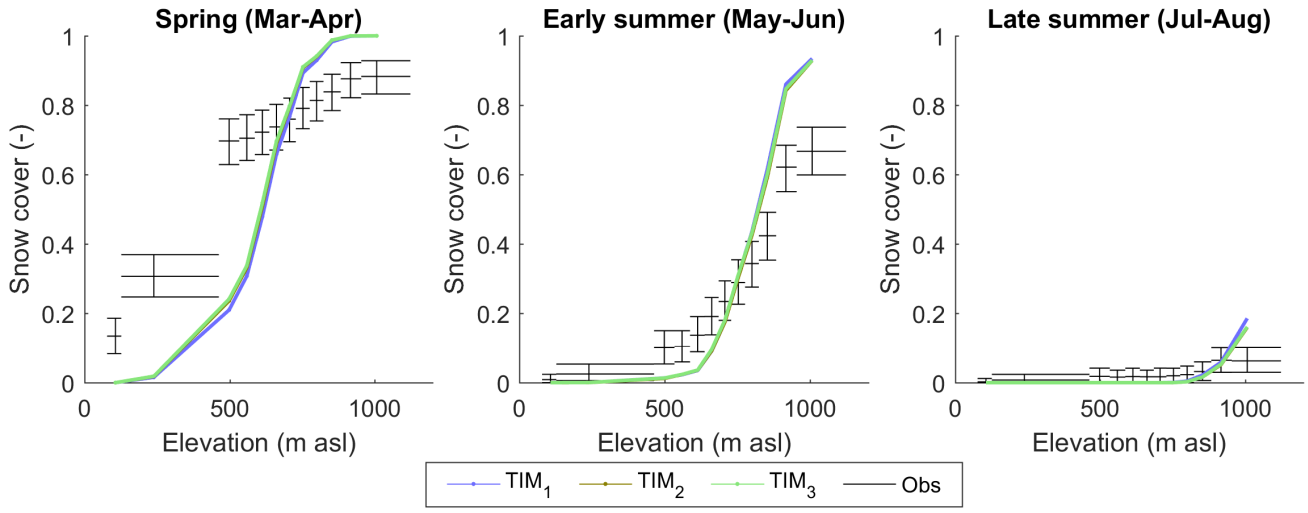

Figure 12. Simulated seasonal snow distribution curves when using the three melt model structures.

TIM_2 model structure is substantially more acceptable than when using the other melt model structures although it should be noted that, given that flows are very low here, the absolute error is less than 0.2 m^3 s^{-1}. ~~For the mean May river flow, the simulation using TIM_3 is less acceptable than when using the other melt model structures. This is interesting, as May coincides with the beginning of the main melt season and therefore this could be some indication of an inability to capture this initialisation properly.~~ A comparison of the simulated ice melt during May 2013 reveals that the TIM_3 structure simulates the highest ice melt of all three melt model structures (Fig. 13a) which results in a positively-biased river flow time-series (see Fig. 13b). Note, the full input/output time-series over the observation period can be found in Appendix D. ~~It has already been shown that the simulated snow coverage signatures are almost identical when using the three melt model structures and as such the contrast in simulated ice melt cannot be directly attributed to the dynamic snow albedo component of TIM_3. A~~ Furthermore, a comparison of the simulated ice melt time-series over 2013 with a monthly moving-average filter ~~reveals~~ demonstrates that the positive melt bias from TIM_3 extends between April and June (Fig. 14b) which corresponds to the period where temperatures are relatively low, but where incoming solar radiation is relatively high (see Fig. 14a). ~~As such, it is the additive form of the TIM_3 melt equation and the subsequent increased influence of solar radiation on melt which induces the bias in flow simulations in the early melt season.~~

Of the remaining river discharge signatures, only a handful show any substantial difference when switching between the melt model structures including the mean April and August discharge and the two 'flashiness' signatures: the integral time and the rising limb density. However, the differences here are very small. For the 'high slope' signature, which characterises the variability of high flow river flows, the simulation using the TIM_1 melt model structure is able to capture it within the LOA, while the simulations using the TIM_2 and TIM_3 model structures both show a negative bias suggesting they underestimate high

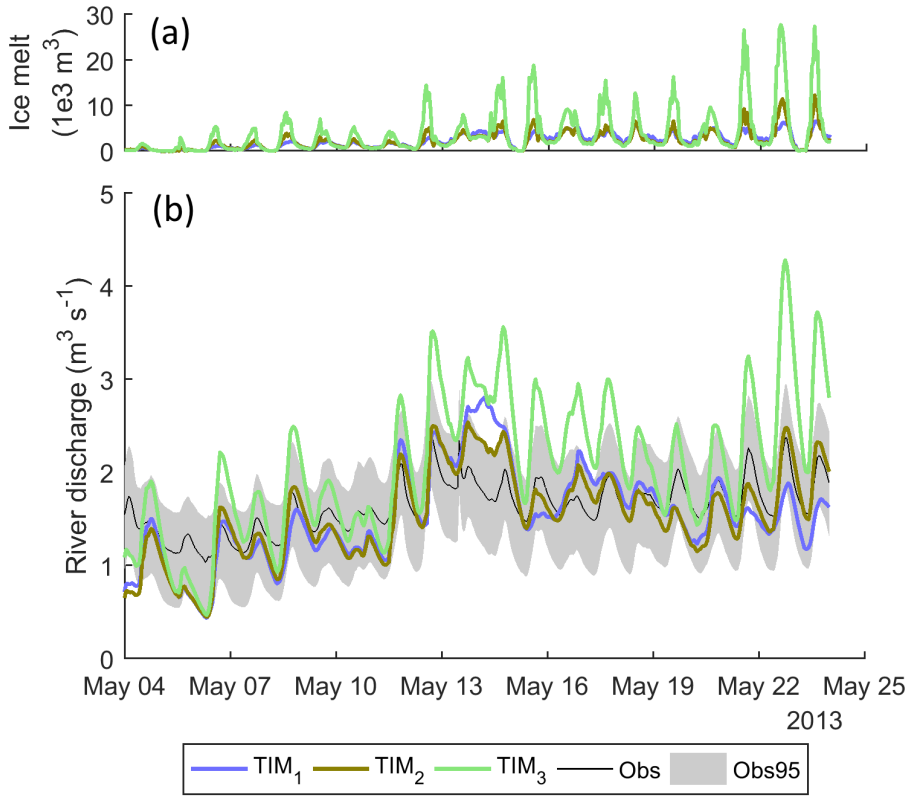

Figure 13. Mean simulated hourly ice melt (a) and river discharge (b) during May 2013 using the top 1% of models from the three melt model structures in combination with the ROR_1 runoff routing model structure.

flow variability. ~~The reason for this is not clear, but it does indicate that the simpler TIM_1 melt model structure is better suited for capturing this signature.~~

3.3 Acceptability of runoff-routing model structures

To evaluate the runoff-routing model structures, acceptability scores have been calculated for the river discharge signatures only as these structures do not influence ice melt or snow coverage (Fig. 15). To ensure fair comparison between the different structures, all scores have been obtained using the simplest TIM_1 melt model structure in the GHM.

It was noted previously, that all melt model structures used in combination with ROR_1 resulted in positively-biased January and February river flows. It could be that including a more complex non-linear runoff-routing model structure in the GHM could help to mitigate this bias. Indeed, the calibrated simulations do show a substantial reduction in positive bias for the mean February flows when using ROR_2 and ROR_3, however the simulations are still unacceptable. Furthermore, for the mean January river flow there is no substantial change in acceptability score. This indicates that the runoff-routing representation is also not the reason for this overestimation of flows at the beginning of the year. To investigate this positive bias further,

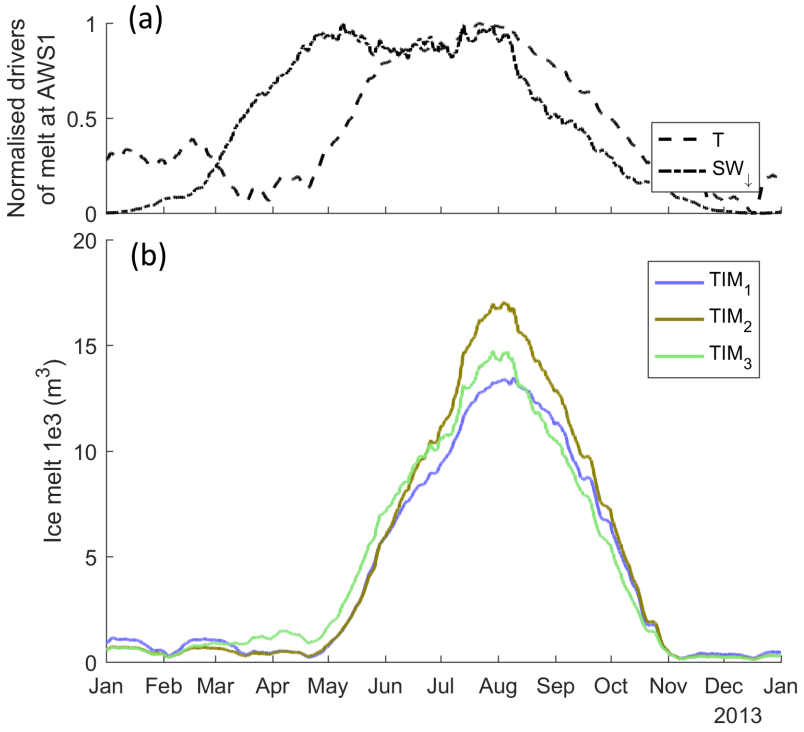

Figure 14. Normalised temperature and incident solar radiation (a) and simulated ice melt from the three calibrated ice melt model structures (b) for the year 2013. All time-series use a monthly moving average filter.

Fig. 16c shows the simulated time-series from the calibrated models using TIM_1 in combination with ROR_1, ROR_2 and ROR_3 for January and February 2013. Figure 16a shows that melt is an insignificant input during these winter months (green line). Rather it is rainfall (black dash) that dominates the runoff input and this results in two pronounced peaks in the simulated river discharge time-series. The different behaviour of the simulations using the three runoff-routing model structures is much more obvious during the rainfall-runoff events. The simulation using the ROR_1 structure is noticeably more flashy in response to the rainfall and overestimates the peak flows while the ROR_2 and ROR_3 simulations, which include additional, more diffusive representations of the flow of water through snow and firn, result in peak flows that are closer to the observed, but with a recession that is too shallow. Regardless of these deficiencies, however, all result in an almost identical positive bias as shown by the cumulative flow in Fig. 16b. ~~The more complex routing structures, therefore, appear to offer no additional capabilities to correct the positive monthly flow biases.~~

There are however differences when assessing other aspects of the river discharge time-series, particularly in the signatures relating to high flows. In Fig. 15, it can be seen that while the simulation using the ROR_1 routing model structure is able to capture all of the high flow signatures simultaneously, the ROR_2 and ROR_3 structures show an unacceptable negative bias for these signatures indicating underestimation of high flow magnitude and variability. To evaluate this in more detail, Fig. 16f

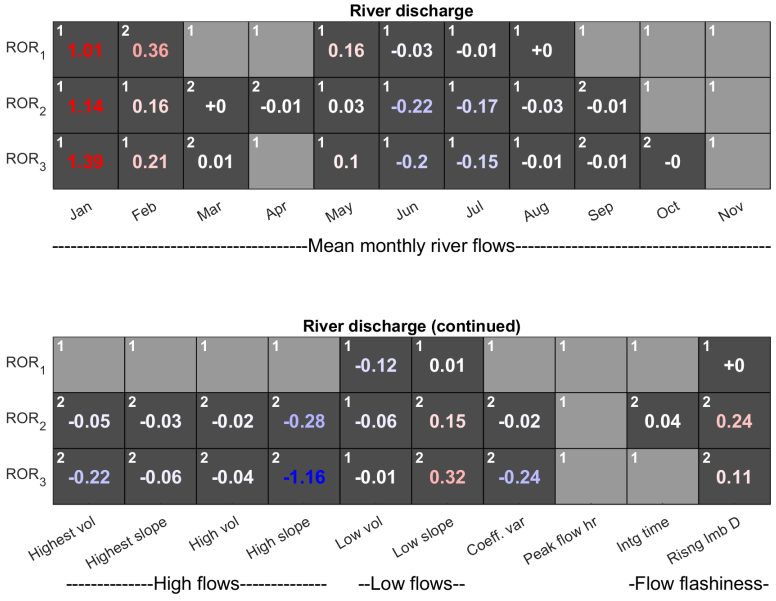

Figure 15. Acceptability scores obtained after calibrating the GHM using the three runoff-routing model structures in combination with the TIM_1 melt model structure. Light grey boxes indicate acceptable simulations ($s = 0$) and numbered, dark-grey boxes indicate unacceptable simulations coloured blue and red to indicate negative and positive bias respectively. Note, all acceptability scores are rounded to two decimal places. Those non-zero scores that round to zero are accompanied by +/- to indicate sign of score. White numbers in top left of each box indicate relative ranking where acceptability scores are substantially different between the GHM configurations.

shows the simulated time-series for the highest recorded river flow event during October 2014. Here, the flashier and more responsive ROR_1 structure achieves the closest fit to the observed peak flow and within the uncertainty bounds while the more diffusive, ROR_2 and ROR_3 structures underestimate the peak flow. Note they also underestimate the overall river flow variability as indicated by the coefficient of variation signature. ~~It appears, therefore, that the diffusive behaviour of the ROR_2 and ROR_3 runoff-routing structures which is advantageous for capturing peak flows in winter, results in underestimation of peak flows at the end of the melt season and an underestimation of overall river flow variability.~~

3.4 Consistency of melt model structures

The results so far have highlighted some inconsistencies in the GHM configurations using the melt and runoff-routing model structures where they are unable to reconcile some combinations of signatures simultaneously. This is important as those inconsistencies could help to further diagnose structural deficiencies in the different model structures. To investigate this, consistency scores have been calculated between pairs of the 33 signatures for each GHM configuration. A model can be deemed consistent across a pair of signatures if it is able to capture both within their LOA simultaneously. The consistency

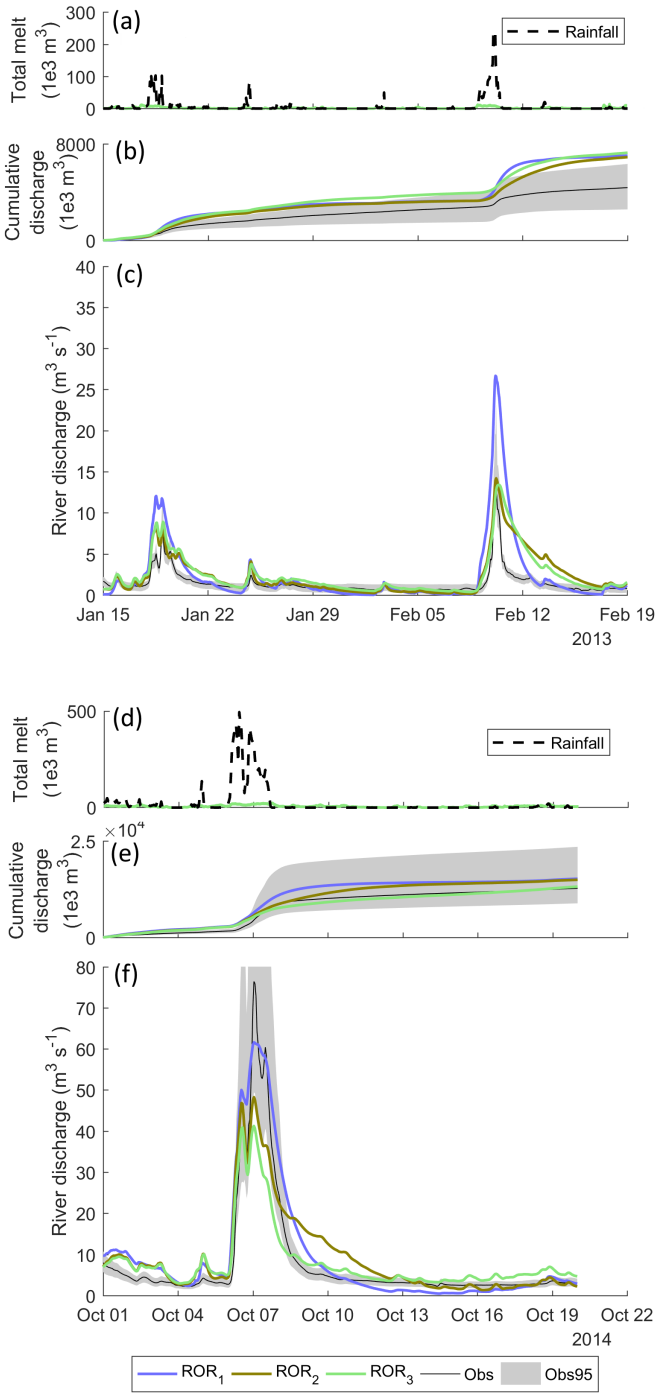

Figure 16. Simulation time-series using the three different runoff-routing model structures in combination with the TIM_1 melt model structure including simulated total melt and rainfall (top), cumulative river discharge (middle) and river discharge time-series (bottom) for January and February 2013 (a,b,c) and the October 2014 flood (d,e,f).

scores are therefore calculated as the minimum sum of the two acceptability scores between a pair of signatures across the 5000 calibration runs for each GHM configuration.

Figure 17 shows the average consistency scores calculated across the signatures for each attribute of ice melt, snow coverage and river discharge using the three melt model structures in combination with the ROR_1 runoff-routing structure. The top panel shows the consistency scores when using the simplest TIM_1 melt model structure. The regions in red highlight the areas where the GHM is inconsistent. The first striking observation is the red band along the upper catchment snow coverage attribute. It has already been demonstrated that the simulations using the TIM_1 structure cannot reconcile the upper-catchment snow coverage with the remaining snow coverage signatures. This further demonstrates that when using the TIM_1 structure, the GHM cannot reconcile the upper-catchment snow coverage with any of the other attributes.

The largest inconsistency score obtained was between the short term, seasonal melt on the glacier tongue and long-term total glacier volume change. ~~This is further evidence that the TIM_1 melt model structure is not able to reconcile the melt signatures over the differing temporal and spatial scales. This raises the question of where this inconsistency stems from.~~ It should be noted that the seasonal melt signatures show a small inconsistency with the lower-catchment snow coverage and a larger inconsistency with the upper-catchment snow coverage. The total glacier volume change signature, however, is also inconsistent with the monthly flow and low flow signatures indicating that it is the long-term glacier wide mass balance that the model is getting wrong.

The use of the TIM_2 model structure which includes topographic effects goes some way to reducing most of the inconsistencies shown using the TIM_1 model structure (Fig. 17 middle panel). However, all but one of the inconsistencies (between lower-catchment snow coverage and seasonal melt) remain, indicating that the use of the TIM_2 melt model structure only provides a small improvement in model consistency.

Using the TIM_3 model structure also helps to improve model consistency, particularly those associated with the upper snow coverage, but surprisingly it also introduces new inconsistencies in relation to the lower-catchment snow coverage, where the model is not able to reconcile these signatures with any of the other attributes. ~~This is interesting, because although the simulated snow distribution curves from each melt model structure were approximately identical, the alteration of the melt equation and inclusion of dynamic snow albedo reduces overall model consistency.~~

3.5 Consistency of runoff-routing model structures

Consistency scores have also been calculated for each pair of river discharge signatures (Fig. 18) using the three runoff-routing structures in combination with the TIM_1 melt model structure. The simulations using the ROR_1 structure (top panel) and next simplest ROR_2 structure (middle panel) show a very similar pattern of model inconsistencies. Firstly, both sets of simulations do not capture the relatively low flows in February and the relatively high flows in July and August simultaneously. This corroborates the findings from the acceptability analysis which revealed a tendency for the model structures to overestimate low flows in the winter and underestimate high flows in the summer and autumn, particularly with relation to rainfall-induced high flows. Interestingly though, the seasonal flow inconsistency is centred on February and there are not inconsistencies for the other low flow months from January to April. This provides further evidence that it is particularly the rainfall-induced flows that

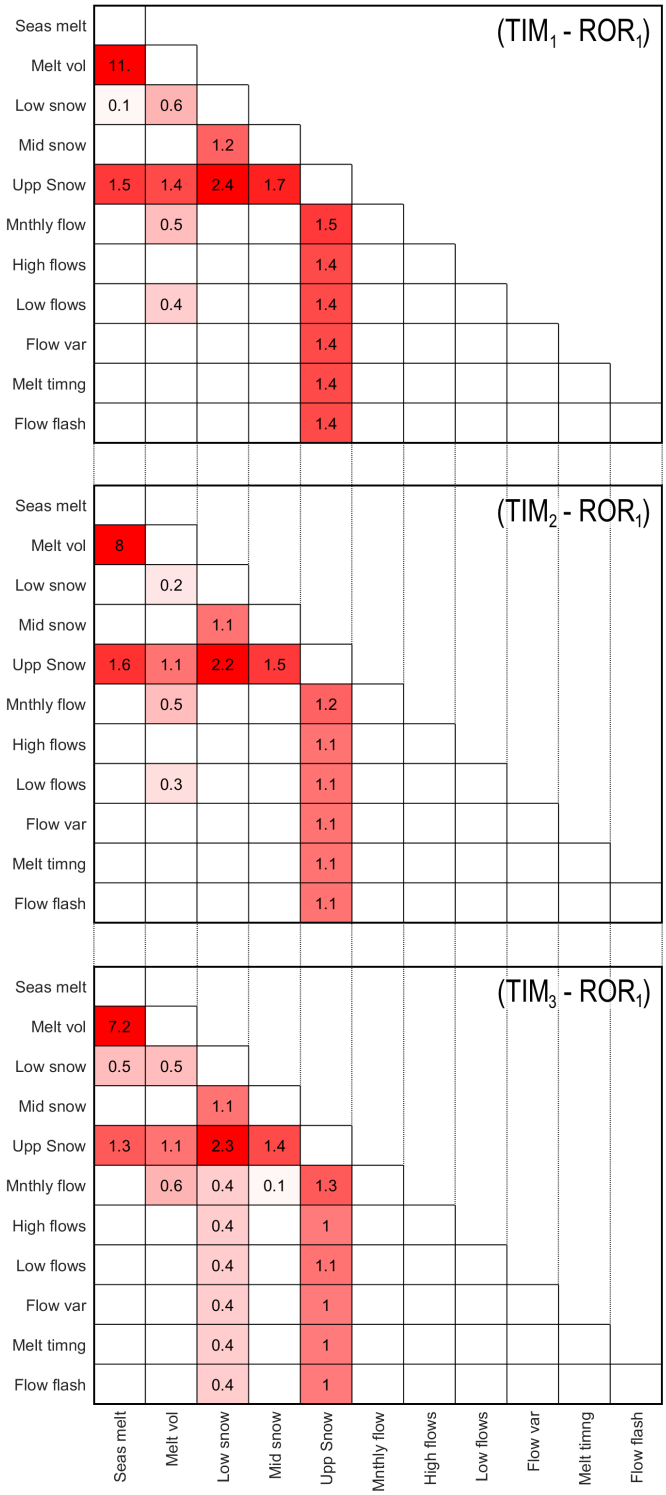

Figure 17. Average consistency scores between attributes using the three melt model structures in combination with the ROR_1 runoff-routing structure. Scores of < 0.1 have not been reported.

the model is not able to capture effectively. In fact, February has some of the highest flows in the record of winter flows induced by large rainfall events (see average flow signatures in Table 2). This suggests this could be the reason that the inconsistencies between winter and summer flows are centred around these months. The inclusion of additional flow pathways in the routing routine only enhances these inconsistencies, particularly when using the ROR_3 model structure where the inconsistencies extend into June (bottom panel).

The ROR_1 simulations show inconsistencies between the February flows and low flow variability as indicated by the low slope signature. The reason for this is not clear, but interestingly, the inclusion of an extra, more diffuse, flow pathway in the ROR_2 model appears to remedy this, suggesting that there is some non-linear behaviour that the ROR_1 model structure cannot capture. However, it comes at the cost of inducing an extra inconsistency between the mean flows in January and the overall flow variability as indicated by the coefficient of variation. This new inconsistency is amplified when using the ROR_3 structure. ~~This further corroborates the findings from the acceptability scores where the more complex runoff-routing model structures cannot capture the observed flow variability adequately.~~

Interestingly, the consistency scores when using the ROR_1 and ROR_2 structures are relatively similar, with each configuration demonstrating inconsistencies between four and five pairs of river discharge signatures respectively. In contrast, using the most complex ROR_3 structure introduces a number of new inconsistencies with a total of 12 inconsistent pairs of simulated river discharge signatures. These new inconsistencies are centred around the mean monthly flow signatures as well as the signatures relating to high and low flow magnitude and variability.

4 Discussion

The first aim of this study was to investigate if a signature-based approach within a LOA framework could be used to diagnose deficiencies in the different melt and runoff-routing model structures. The comprehensive set of signatures provided a powerful method to evaluate the model behaviour. Furthermore, when used within ~~the~~ a LOA framework, it was straightforward to identify those aspects of the glacio-hydrological system that the GHM configurations could not capture. A number of the identified model deficiencies are particularly important in the context of future river flow predictions which will now be discussed.

Regardless of the choice of melt model structure, all GHM configurations were able to capture the three signatures of ice melt individually, but none of them could capture all of the signatures simultaneously. The challenge here was to reconcile three signatures that characterise glacier melt over different spatial and temporal scales. This is not a straightforward task, particularly when using temperature index models that lump a number of spatially and temporally variable terms from the full energy balance equation into a handful of calibration parameters which may lack robustness in space and time ~~(MacDougall et al., 2011; Matthews et al., 2015; Gabbi et al., 2014)~~(MacDougall et al., 2011; Gabbi et al., 2014; Matthews et al., 2015). The inclusion of solar and topographic effects in the TIM_2 and TIM_3 melt model structures addressed some of these limitations. Indeed, the inclusion of these in conjunction with the dynamic snow albedo parameterisation returned the most acceptable simulations of the ice melt signatures overall. However, further improvements are required to achieve acceptable model simula-

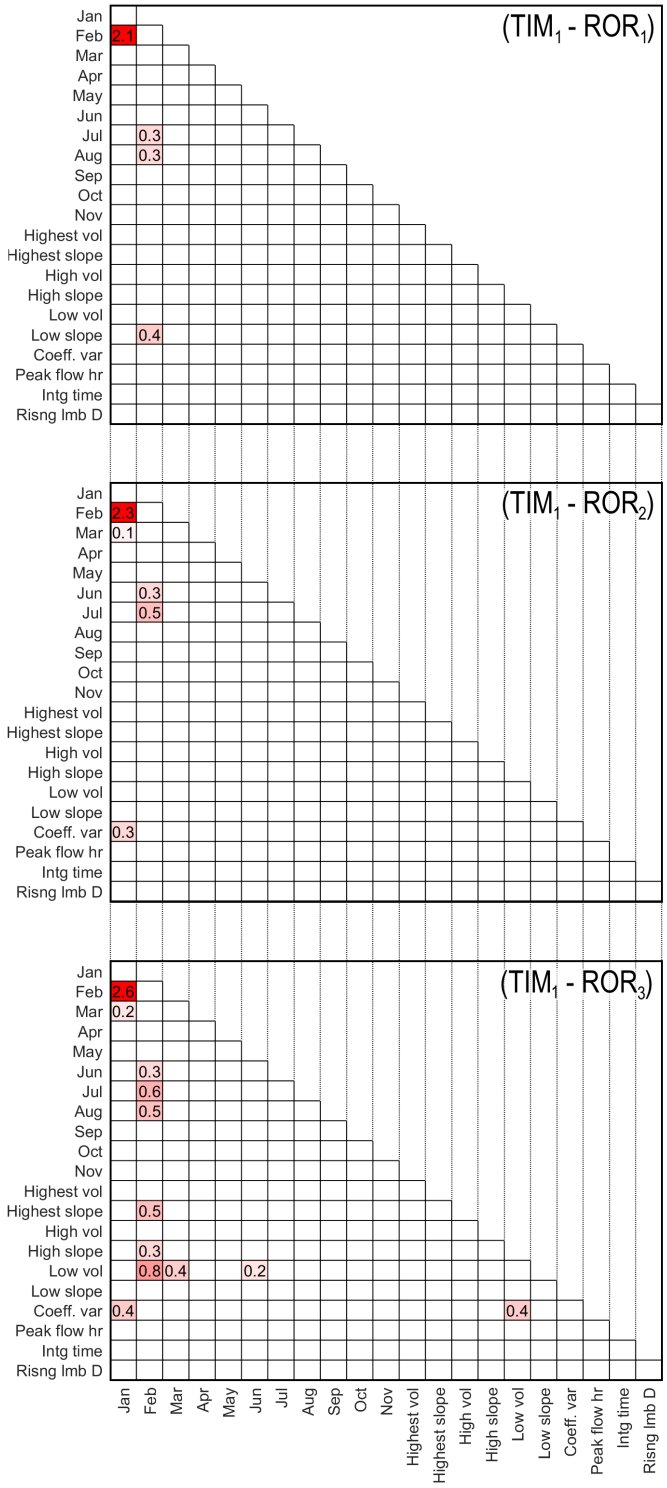

Figure 18. Average consistency scores between river discharge signatures using the three runoff-routing model structures in combination with the TIM_1 melt model structure.

tions that capture all of the ice melt signatures simultaneously. Certainly, one aspect of the glacier which was not accounted for was debris cover at the glacier terminus which could be an important control on point scale and overall ice mass balance. Some TIMs that include representations of debris cover do exist (e.g. Carenzo et al., 2016) and the signature-based LOA approach would provide the ideal framework for evaluating the added-value of further structural modifications like these.

The snow coverage signatures highlighted deficiencies in all of the GHM configurations. None of the prior 45,000 model compositions were able to capture the spring and early-summer snow coverage in the upper catchment and all of the calibrated GHM configurations overestimated snow coverage in the upper catchment whilst underestimating it in the lower catchment. Interestingly, using the most sophisticated TIM_3 structure with the dynamic snow albedo function had almost no effect on the overall acceptability across these signatures, indicating that the melt model formulation was not the primary source of model deficiencies here. Of course, snow coverage simulations are sensitive to other components of the GHM such as the snow redistribution model, which itself, is sensitive to the resolution of the DEM used to parameterise it; a coarser DEM resolution removes peaks and troughs in the land surface which can bring about more complex patterns of snow coverage. Similarly, the glacier volume change signature will be sensitive to the glacier evolution formulation and parameterisation. It is clear, therefore, that while the application of ~~the~~ a LOA framework here has demonstrated the gains that can be made in capturing some signatures through the inclusion of extra model complexity, the apparent insensitivity of the snow coverage signatures to structural modifications indicates that further gains may also be made by investigating other components of the GHM structure within this framework. ~~Indeed, beyond~~ Beyond the structural nature of a GHM, the boundary conditions ~~used~~ may also contribute to model deficiencies. ~~In particular~~For this study, the driving precipitation data ~~, which while~~ was relatively well constrained by observations ~~at the~~ within the catchment during the summer and autumn months of recent years, but there were fewer observations during the winter months and none at all before 2009. Furthermore, while the bias-corrected precipitation time-series was well correlated over a three-day time-step, it was not at an hourly time-step. It's also important to note that precipitation observations were all collected at the bottom of the catchment ~~, are less certain higher up in the catchment where limited observation data are available~~and therefore driving precipitation data at the top of the catchment are less certain. Indeed, one could explain the tendency to overestimate snow coverage higher up in the catchment by a positive bias in the driving precipitation data here. Such a bias could also explain the modelled inconsistencies across ~~the ice melt signatures. However, before making such a conclusion, it~~ signatures that characterise ice melt at different spatio-temporal scales. Furthermore, given the strong coupling between snow, ice and river runoff, deficiencies in capturing the snow and ice signatures could also propagate through the hydrological representation of the catchment. For example, one could imagine how errors in the spatial distribution of snow could perturb the timing of runoff through the catchment given that snow distribution influences the behaviour of the semi-distributed runoff-routing routine employed in the GHM. Such perturbations are likely to impact the ability of the GHM to capture the full range of river discharge signatures. Accordingly, it is important to stress the influence that biases in the driving climate data could have on the model acceptability across the different signatures. Of course, for balance, it should be noted that ~~the~~ regarding precipitation higher up in the catchment, the limited precipitation collected at the summit of Öræfajökull indicate that ~~the~~ mean annual biases in driving precipitation data ~~used in this study are~~

~~approximately correct~~ are small (Guðmundsson, 2000). Even so, the recent melt model comparison by Reveillet et al. (2017) suggest that uncertainties in driving precipitation data can cloud any differences between melt model behaviour.

As a side experiment, we tried increasing the snow melt parameters for the TIM_3 model to see if this would help remove some of the inconsistencies across the ice and snow signatures. It was found that using snow melt parameters equal to or even greater than those used for the ice reduced (improved) the consistency score between the ice melt signatures by more than 50% when using TIM_3. The physical justification of this is of course questionable, but it does highlight the influence that the prior parameter distributions could have on the results presented here. Accordingly, different specifications of the calibration parameter ranges may also help to improve model consistency.

The main deficiencies noted for all of the GHM configurations when compared to the river discharge signatures were an overestimation of the relatively low winter flows in January and February, and the flows at the start of the melt season in May. It was assumed that the addition of extra 'slow' flow pathways in the ROR_2 and ROR_3 runoff-routing structures would help to correct for any deficiencies in capturing the hydrograph seasonality. Instead, the choice of runoff-routing structure had very little influence on these signatures, indicating that longer term storages of water do not have a major control on the seasonality of the hydrograph. This is probably because of the catchment's small size which leads to an instantaneous seasonal response to melt on a monthly timescale. We suggest, however, that for larger catchments, the monthly flow signatures are likely to be more sensitive to the choice of runoff-routing structure. Instead, the simulated mean monthly river flow signatures were more sensitive to the choice of melt model structure, particularly in May at the start of the melt season, which is not surprising given the high degree of glaciation of the river basin. Even so, regardless of the melt model structure employed, none of the GHM configurations could correct for the biases in mean monthly flows indicating that further structural modifications are required. One process that is not represented at all in any of the GHM configurations, but which has shown to be important in for Icelandic glaciers is refreezing of melt water and rainfall (Johannesson et al., 1995). It is estimated that about 7% of total melt in valley glaciers in Iceland refreezes, and therefore, the inclusion of this process could also help to reduce runoff during the colder months of January and February.

In contrast to the monthly river flow signatures, the choice of runoff-routing structure had by far the dominant control on those signatures that are controlled by flows operating on much shorter timescales such as the distribution of flows, flow variability and flashiness. This hierarchy of influence between the melt and runoff-routing model structures has important implications for river discharge prediction uncertainty in glaciated basins. For example, if one were interested in future seasonal water resource availability, they would be most reliant on predictions of mean monthly river flows. The results here indicate that, for this catchment at least, uncertainties in these predictions stem primarily from melt model uncertainty. In contrast if one were interested in future changes in flood frequency, the dominant source of model prediction uncertainty is the runoff-routing approach. Uncertainties in river flow predictions from glacio-hydrological models are therefore dependent on the river flow characteristic of interest.

The results from the simulated river discharge signatures also raised some questions about the added value of introducing extra complexity to conceptual models of glacio-hydrological processes. The most sophisticated TIM_3 melt structure was the most consistent across the ice melt and snow coverage signatures. However, it was also the least acceptable structure for the

mean May river flow signature ~~, an artefact of its additive form .~~ where it showed the highest positive bias. This is interesting, as May coincides with the beginning of the main melt season and therefore this could be some indication of an inability to capture this initialisation properly. It was shown that the simulated snow coverage signatures were almost identical when using the three melt model structures indicating that this deficiency did not stem from the dynamic snow albedo component of TIM_3. Furthermore, May corresponds to the period where temperatures are relatively low, but where incoming solar radiation is relatively high indicating that it is the additive form of the TIM_3 melt equation and the subsequent increased influence of solar radiation on melt which induced the positive bias in flow simulations in the early melt season.

Similarly, the ROR_3 structure, originally proposed as the most realistic conceptual representation of water storage and transmission in the river basin, was the least acceptable model overall across the river discharge signatures. Certainly, the more diffusive behaviour of the ROR_3 runoff-routing structure was advantageous for capturing peak flows during the winter. However, it also resulted in underestimation of peak flows at the end of the melt season and an underestimation of overall river flow variability. These results highlight the need to exercise caution before introducing complexity to conceptual models of glacio-hydrological processes. They also illustrate the importance of testing prior assumptions about the system against other possible model hypotheses, for which ~~the~~ a signature-based LOA framework is ideally suited.

The second aim of this study was to determine if the signature-based evaluation within a LOA framework could be used to constrain the prior population of model structures and parameter sets (compositions) down to a smaller population of acceptable models. The initial discrimination tests showed that all of the signatures have discrimination power, although for two of the snow signatures, none of the 45000 model compositions could capture them. The mean January and May river flow signatures were the best discriminators, individually reducing mean river discharge uncertainty to 60 - 70% of that from the full population of model compositions, although it should be noted that the majority of this reduction stemmed from constraining the acceptable parameter sets rather than the model structures. These results indicate that ~~the~~ a LOA framework could be used to find a population of acceptable model compositions. However, the fact that none of the prior 45000 compositions were able to capture all of the signatures means that this remains to be seen. At a fundamental level, the results indicate that the structural configurations of the GHM employed in this study are simply not good enough to capture the observation data within their observation uncertainty bounds. To address this, one could implement further structural modifications, some of which have been alluded to, until acceptable simulations are obtained. Indeed, a more thorough exploration of a wider parameter space could also yield acceptable model compositions. There are of course other sources of unacceptable behaviour though. Of these, boundary conditions including the ice and watershed boundaries as well as the driving climate data are all contenders. The initial ice geometry was also uncertain but not explicitly accounted for. It is therefore recommended that where possible, future applications of ~~the~~ a LOA framework should incorporate these additional sources of uncertainty, so that more robust conclusions about model appropriateness can be made. Certainly, ~~study~~ it's important to emphasise that these future applications need not adopt the same 33 signatures used in this study. On the contrary, the choice of signatures will always depend somewhat on the availability of data at a given study site as well as the complexity (e.g. spatio-temporal resolution) of the model(s) being interrogated. Indeed, future users should be encouraged to experiment with different signatures (where data permits)

particularly if they wish to focus on other process representations within their GHM. Study sites with good observation data and understanding of data uncertainty would be ideal candidates for these future applications.

5 Conclusions

The signature-based, LOA framework adopted in this study provided a comprehensive evaluation of different GHM melt and runoff-routing model structures. In contrast to traditional model evaluation approaches which rely on one or several global summary statistics, the adoption of multiple signatures helped to identify those aspects of the glacio-hydrological system that a particular model could or could not capture and the added value of introducing additional complexities to simplified process models. When evaluated against individual signatures, the more complex model formulations did improve model simulations in some cases. However, they were not necessarily more consistent across the full range of signatures, emphasising the need to exercise caution and properly evaluate if additional complexities are justified. The often conflicting acceptability scores across the signatures highlights the difficulty and inherent uncertainty in model structure selection. It is clear, therefore, that future glaciological and hydrological projection studies that use simplified model structures should take account of these uncertainties, although to date these have rarely been considered. For future river flow predictions in glaciated basins it is likely that the source of model uncertainty depends on the particular river flow characteristic of interest. We found evidence that a hierarchy of influence exists between the melt and runoff-routing model structures across the range of river discharge signatures.

An additional advantage of adopting ~~the~~ a LOA framework is that it provides objective criterion for accepting or rejecting particular model structures and parameterisations. While all ~~, but two ,~~ but two of the signatures demonstrated discrimination power, none of the 45,000 different model compositions tested in this study were able to capture them within their LOA simultaneously. Therefore, it remains to be seen if the framework can be used in this way, although we suggest that applications that go beyond examining the melt and runoff-routing structural uncertainties may prove more fruitful in obtaining a behavioural population of models. These should consider other uncertainties including those associated with snow redistribution, glacier evolution and model boundary conditions. We would therefore encourage future studies, particularly where a broad range of observation data covering different aspects of the glacio-hydrological system, to move away from using traditional global summary statistics for model evaluation and adopt a multi-metric approach within a LOA framework so that their simplified process hypotheses can be rigorously tested, and structural uncertainty better understood.

Code and data availability. For persons interested in applying a similar signature-based LOA approach for model evaluation, we would encourage them to contact the authorship who are open to providing advice and sharing data and code where possible.

Appendix A: Glacio-hydrological model

A1 Soil infiltration and evapotranspiration

The semi-vegetated nature of the catchment coupled with the relatively cool temperatures year-round mean that evapotranspiration is generally low (Einarsson, 1972). Even so, to satisfy the water balance, an explicit representation of the soil zone for model nodes that are not ice or snow-covered was included using the method developed by Griffiths et al. (2006) which has been successfully applied to temperate regions in the past ~~(?Sorensen et al., 2014)~~ (Mackay et al., 2014; Sorensen et al., 2014) and is based on the well established UN Food and Agricultural Organisation soil water balance method (Allen et al., 1998). For each bare ground node, the soil is represented as a finite storage reservoir with a soil water capacity, termed the total available water, TAW [L], which defines the maximum volume of water available to plants for evapotranspiration after the soil has drained to its field capacity and can be defined from lookup tables with basic information on vegetation and soil information (Allen et al., 1998). This was parametrised using the 'Talus' soil class and 'semi-vegetated' land surface class giving an average TAW value of 7 mm. Soil storage is replenished by infiltration from rainfall and melting of residual snow overlying the bare ground and is depleted by evapotranspiration giving a soil water balance:

$$\frac{\Delta S_{soil}}{\Delta t} = I - ET \tag{A1}$$

where S_{soil} [L] is the soil water storage, t is time, I [LT^{-1}] is the infiltration rate and ET [LT^{-1}] is the evapotranspiration rate. Because measured ET is rarely available, Griffiths et al. (2006) propose using the potential evapotranspiration rate, ET_0, instead which defines the evapotranspiration rate from a reference grass covered wet soil (see Appendix A2 for calculation of ET_0). Using ET_0 as the maximum possible evapotranspiration rate, they define a separate function which accounts for the fact that as the soil becomes drier, plants find it more difficult to extract moisture from the soil matrix, and therefore ET is typically less than ET_0. While this is conceptually sound, it was decided not to include this function and instead assume that $ET = ET_0$. There are three reasons for doing this. Firstly, because the inclusion of this function requires an additional parameter which is uncertain and must be calibrated. Secondly because ET is a relatively small component of the overall water balance in this catchment and it was not the aim of this study to investigate this aspect of the catchment hydrology. Thirdly, because previous studies have shown that this parameter (and therefore the behaviour of this function) is relatively insensitive and unidentifiable ~~(?)~~(Mackay et al., 2014).

In the original formulation by Griffiths et al. (2006), any excess soil water (i.e. when $S_{soil} > TAW$) is distributed between overland flow and groundwater recharge pathways. They use a fixed baseflow index (BFI) parameter which defines the proportion of soil water excess that recharges the groundwater. Given the nature of the Virkisá river basin (thin soils overlying impermeable bedrock), it was assumed that soil water migrates to the river outlet via relatively fast, overland flow pathways only and so the BFI parameter was set to zero.

A2 Potential evapotranspiration

Potential evapotranspiration can be calculated from measured meteorological data, most simply as a linear function of measured temperature (e.g. Blaney and Morin, 1942), or where measurements of windspeed, air pressure and solar radiation exist, the full Penmen-Monteith combination equation can be solved. Given that these additional variables are measured at AWS1 from 2009, the combination equation as defined by Allen et al. (1998) was used to calculate hourly potential evapotranspiration over this period:

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} u(e_s - e_a)}{\Delta + \gamma(1 + 0.34u)} h \quad \text{(A2)}$$

where ET_0 is the daily average potential evapotranspiration rate (mm d^{-1}), R_n is the net radiation (MJ m^{-2} d^{-1}), G is the soil heat flux (MJ m^{-2} d^{-1}), e_s and e_a are the saturation and actual vapour pressure respectively (kPa), Δ is the rate of change of the saturation vapour pressure with temperature (kPa $^oC^{-1}$), γ is the psychrometric constant (kPa $^oC^{-1}$), u is the wind speed (m s^{-1}) and T is the mean daily ambient air temperature (^{o}C).

Prior to 2009, the viability of using T as a proxy for ET_0 in a linear regression model framework like Blaney and Morin (1942) was investigated. Similarly, incident solar radiation was also used as the independent variable for this model. In fact, the best fit was achieved using both variables in a multiple linear regression model which was able to explain 66% of the ET_0 variance (Fig. A1). This model was used to distribute ET_0 in space and time using the driving temperature and incident solar radiation data.

A3 Glacier geometry evolution

The empirical Δ-h parametrisation (Huss et al., 2010) requires the availability of at least two digital elevation models of the glacier separated in time. The difference between the two is used to define the Δ-h polynomial which has the form:

$$\Delta h = (h_r + a)^\gamma + b(h_r + a) + c \quad \text{(A3)}$$

where Δh is the normalised surface elevation, h_r is the normalised elevation range and a, b, γ and c are fitted parameters. For this study, the two digital elevation models from 1988 and 2011 were used to define this relationship. Figure A2 (top) shows the raw change data against the 1988 ice elevation. It was decided that the data at the very front of the glacier should not be used as here the ice has completely melted and as such the bedrock beneath skews the raw change data. Figure A2 (bottom) shows the fitted Δh model to the normalised mean elevation change curve. Following Huss et al. (2010), the glacier geometry is updated each year by distributing the net glacier mass balance across the glacier according to this relationship.

Appendix B: Temperature lapse rates

In order to investigate seasonal variations in lapse rate, the temperature gradient between the lowest (AWS1) and highest (AWS4) weather stations in the Virkisá river basin were analysed. The results showed a remarkable degree of variation in

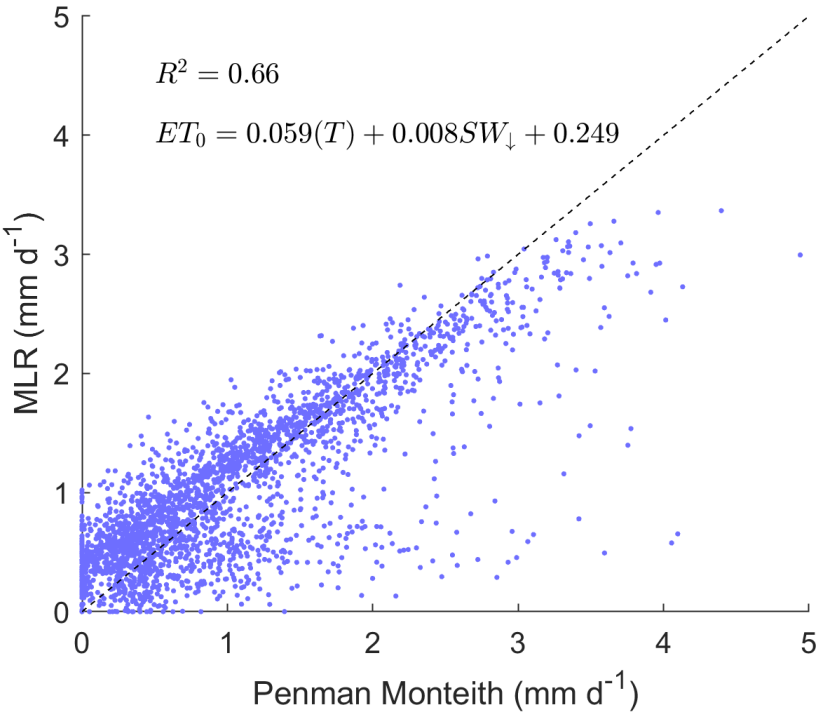

Figure A1. Multiple linear regression model used to convert ambient air temperature and incoming solar radiation into potential evapotranspiration.

hourly average lapse rate between the months of the year (white lines in Fig. B1). During the winter months between November and February, the lapse rate is a relatively stable -5 °C km^{-1} throughout the day. In contrast, between March and October there is a pronounced diurnal variation in lapse rate where it is strongest in the late afternoon/early evening. The heat maps in Fig. B1 represent the frequency distribution of wind direction for each month and show that the development of the strongest lapse rates in the afternoon correspond with a break up of the prevailing northeast winds that flow down from the summit of Öræfajökull and a switch to winds from the southwest. Petersen and Pellicciotti (2011) found a similar phenomenon on the Juncal Norte Glacier in the semi-arid Chilean Andes. They attributed the shallow temperature gradient in the morning with katabatic winds flowing down glacier which serve to cool the air over the glacier and weaken the lapse rate. In the afternoon, they showed that a breaking up of this layer by valley winds served to increase the temperature gradient by warming the air over the lower glacier. This suggests that winds flowing down from the Öræfajökull summit in the warmer months could serve to cool near surface air temperatures over the ice, thereby retarding ice melt. To account for this phenomenon, Petersen and Pellicciotti (2011) suggest adopting the Shea and Moore (2010) model to correct on-ice temperatures relative to ambient off-ice weather station measurements. Shea and Moore (2010) found that for three glaciers on the southern Coast Mountains of British Columbia,

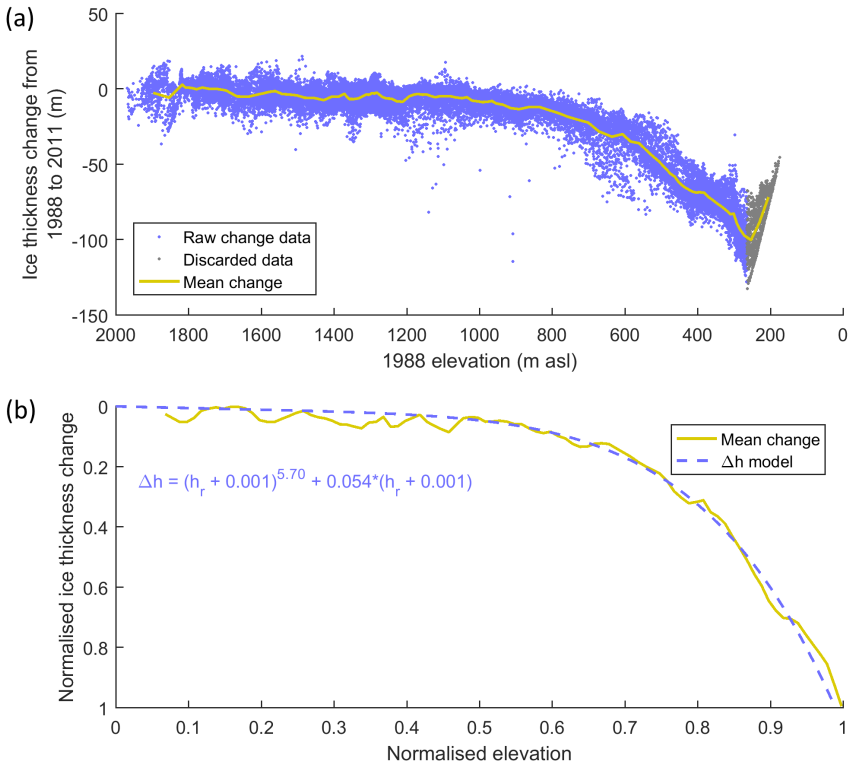

Figure A2. Raw elevation change data from 1988 and 2011 ice DEMs (a) and fitted Δh model to normalised mean elevation change curve following Huss et al. (2010) (b).

Canada, there was a threshold in ambient off-ice air temperature, above which the winds flowing over the glacier served to cool the near-surface on-ice air. They suggest this temperature lies somewhere between 4 and 8 °C, but is likely to be site specific.

To investigate if such a threshold exists on the Virkisjökull glacier, five Gemini Tinytag Aquatic 2 temperature loggers were deployed across the glacier at elevations ranging from 150 - 400 m asl. Each logger was secured at 1.5 m above the ice in a white PVC radiation shield attached to a tripod (Fig. B2). The sensors were deployed for 7 days in late August 2016 and then for a further 7 days in early March 2017 to represent summer and winter on-ice temperatures respectively. The loggers were synchronised in time with the AWS weather stations and set to measure temperature every 15 minutes. This allowed for the direct comparison of on and off-ice near surface temperatures.

Figure B3 shows the synchronised on and off-ice temperatures from all of the measurements taken in winter (blue dots) and summer (yellow dots). The off-ice temperatures were derived assuming a linear lapse-rate between AWS1 and AWS3 as these are situated at elevations similar to the Tinytag temperature loggers. The results show that there is a temperature threshold above which on-ice temperature falls below off-ice temperature which was estimated to be 5.27 °C. Following ~~Petersen and Pellicciotti (2011); Shea and Moore (2010); Ragettli et al. (2014)~~Shea and Moore (2010); Petersen and Pellicciotti (2011); R , this cooling effect was interpreted as being due to northeast winds which bring cooler air from above over the tongue of the

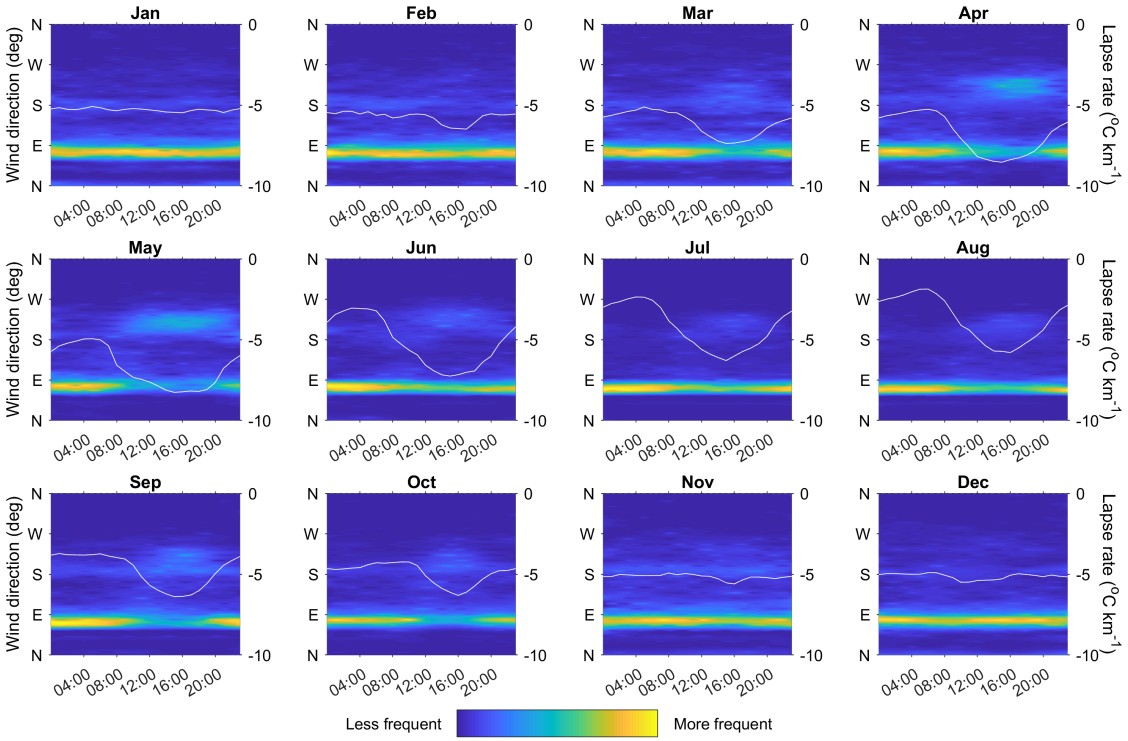

Figure B1. Monthly average hourly temperature lapse rates (white lines, right hand axis) derived from AWS1 and AWS4 temperature time-series overlying heat maps which represent the frequency distribution of hourly wind direction data from AWS4 (left hand axis).

glacier, thereby cooling the on-ice air temperature and the piecewise function derived from Fig. B3 was employed to correct temperatures on the ice during the warmer months when ambient air temperatures exceed this threshold:

$$T_{on} = \begin{cases} T_{off} & T_{off} \leq 5.27 \\ 0.74 \cdot T_{off} + 1.38 & T_{off} > 5.27 \end{cases} \quad \text{(B1)}$$

where T_{on} and T_{off} are the on and off-ice near-surface air temperature (°C).

Appendix C: Calibration parameters

Table C1 lists all of the calibration parameters for the melt and runoff-routing model structures which were randomly perturbed during the GHM calibration procedure.

Appendix D: GHM input/output time-series

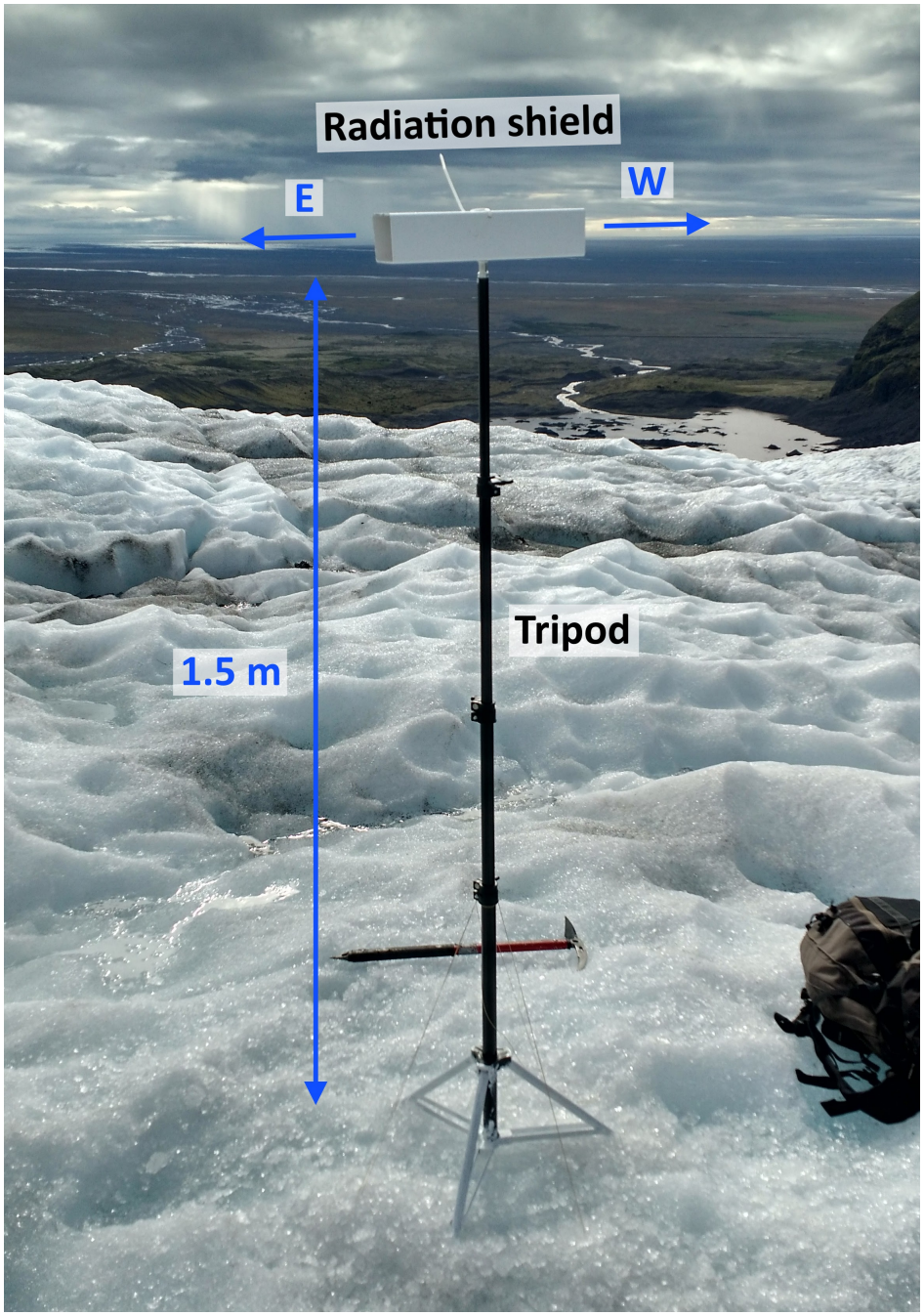

Figure B2. Example of Gemini TinyTag housing used for measuring on-ice temperature at one location on ice.

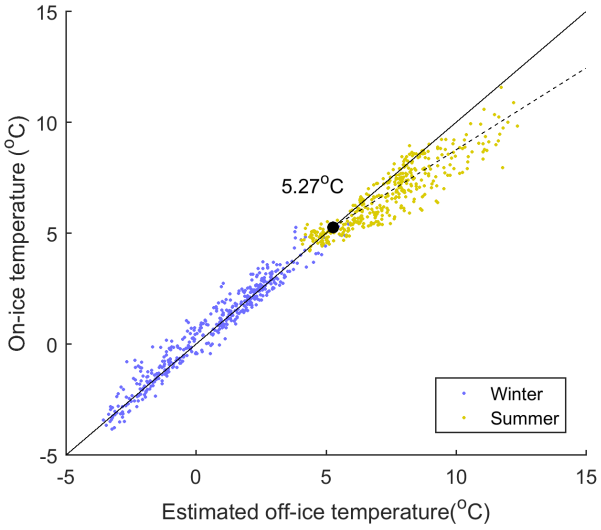

Figure B3. Derived temperature threshold where on-ice temperature is cooler than the ambient off-ice temperature using Shea and Moore (2010) model.

Table C1. Calibration parameters for the melt and runoff-routing model structures.

Structure	Parameter	Calibration range	Units
TIM_1	a_{ice}	2.0e-4 - 7.0e-4	m we $^oC^{-1}$ hr^{-1}
	$a_{snow/firn}$	4.0e-7 - 2.0e-4	m we $^oC^{-1}$ hr^{-1}
TIM_2	a_{ice}	2.0e-4 - 7.0e-4	m we $^oC^{-1}$ hr^{-1}
	$a_{snow/firn}$	4.0e-7 - 2.0e-4	m we $^oC^{-1}$ hr^{-1}
	b_{ice}	4.0e-7 - 2.0e-6	m^3 we W^{-1} $^oC^{-1}$ hr^{-1}
	$b_{snow/firn}$	4.0e-8 - 4.0e-7	m^3 we W^{-1} $^oC^{-1}$ hr^{-1}
TIM_3	a_{ice}	1.5e-4 - 3.0e-4	m we $^oC^{-1}$ hr^{-1}
	$a_{snow/firn}$	6.0e-5 - 2.0e-4	m we $^oC^{-1}$ hr^{-1}
	b_{ice}	1.0e-5 - 8.0e-5	m^3 we W^{-1} hr^{-1}
	$b_{snow/firn}$	2.0e-7 - 4.0e-6	m^3 we W^{-1} hr^{-1}
	p_2	0.01 - 0.4	
ROR_1	k	1 - 30	hr
	n	1 - 5	
ROR_2	$k_{ice/soil}$	0.1 - 5	hr
	$k_{snow/firn}$	20 - 100	hr
	$n_{ice/soil}$	1 - 5	
	$n_{ice/snow}$	1 - 5	
ROR_3	k_{soil}	0.1 - 5	hr
	k_{ice}	0.1 - 5	hr
	k_{snow}	10 - 50	hr
	k_{firn}	50 - 300	hr
	n_{soil}	1 - 5	
	n_{ice}	1 - 5	
	n_{snow}	1 - 5	
	n_{soil}	1 - 5	

Figure D1 shows the complete GHM input and output time-series over the period with observed river discharge data. These include the watershed total precipitation, watershed average temperature and incident solar radiation data used to drive the GHM as well as the simulated watershed total snow melt, ice melt and river discharge using the TIM_1, TIM_2 and TIM_3 melt model structures in conjunction with the simplest ROR_1 runoff-routing structure. Figure D2 shows the same set of plots when using the ROR_1, ROR_2 and ROR_3 runoff-routing model structures in conjunction with the simplest TIM_1 melt model structure.

Author contributions. JDM ran all model experiments and conducted the analysis of results. He also led the writing of this manuscript. All co-authors contributed to formulation and discussion of methods used as well as writing of manuscript.

Competing interests. The authors declare they have no competing interests.

Acknowledgements. This work was supported by a NERC studentship awarded to JDM via the Central England NERC Training Alliance (CENTA). The authors' acknowledge the support of Joaquin Maria Munoz Cobo Belart (University of Iceland) for providing the 1988 ice DEM of Öræfajökull as well as Dr Andrew Black (University of Dundee) and Lee Jones (British Geological Survey) for providing the river discharge and terrestrial LIDAR data used in this study. We would also like to acknowledge the useful discussions with Prof Jim Freer and Dr Gemma Coxon (University of Bristol) regarding the implementation of the limits of acceptability framework. Finally, we acknowledge the assistance given by Heiko Buxel (British Geological Survey) for collecting the on-ice temperature measurements. JDM, CRJ and JE publish with permission of the Executive Director of the British Geological Survey.

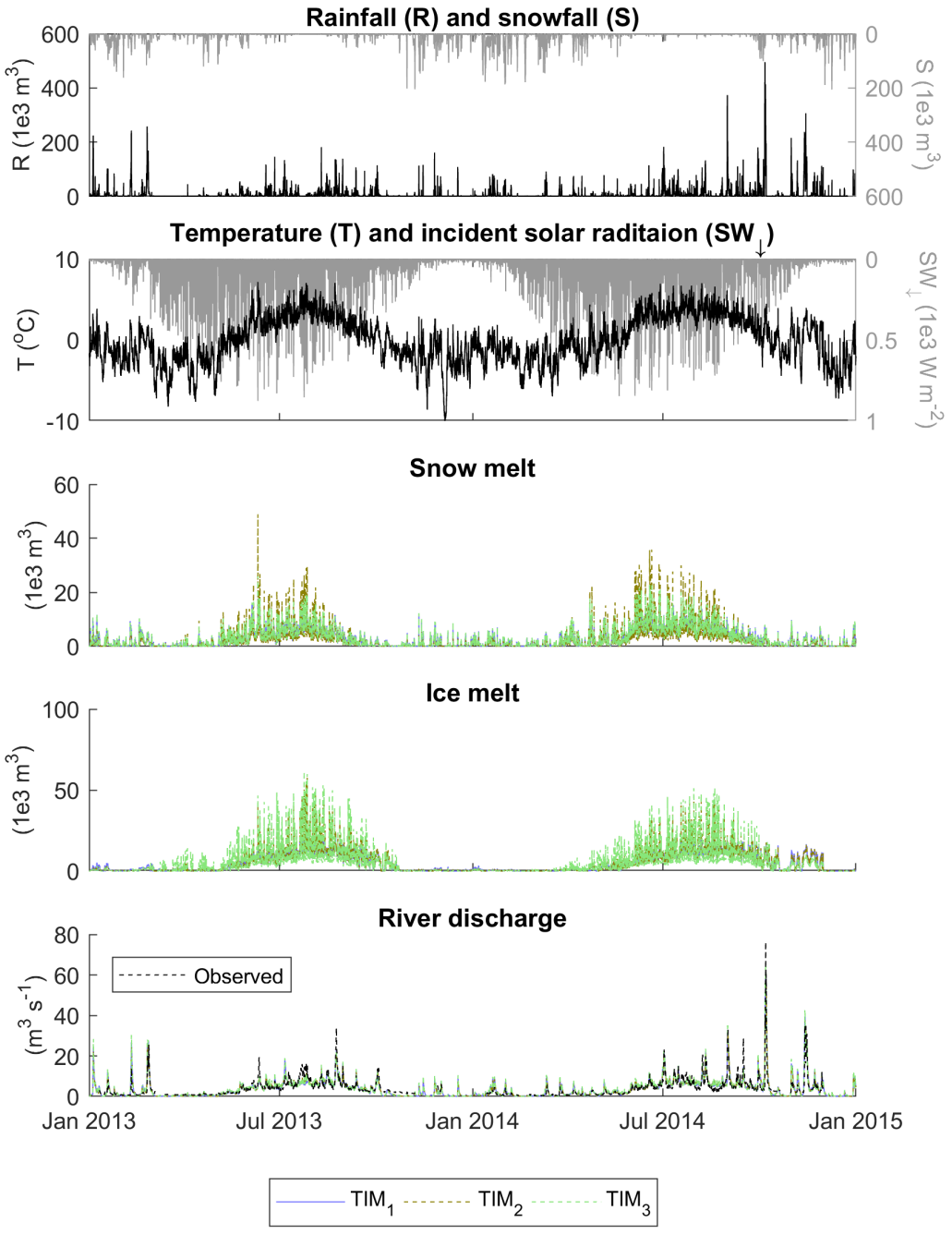

Figure D1. Time-series of driving precipitation, temperature and incident solar radiation data and simulated snow melt, ice melt and river discharge using the TIM_1, TIM_2 and TIM_3 melt model structures in conjunction with the simplest ROR_1 runoff-routing structure. Note, the proportion of rainfall and snowfall is an output from the GHM which is approximately equal across the different configurations. Ice melt includes melt of bare ice and the firn.

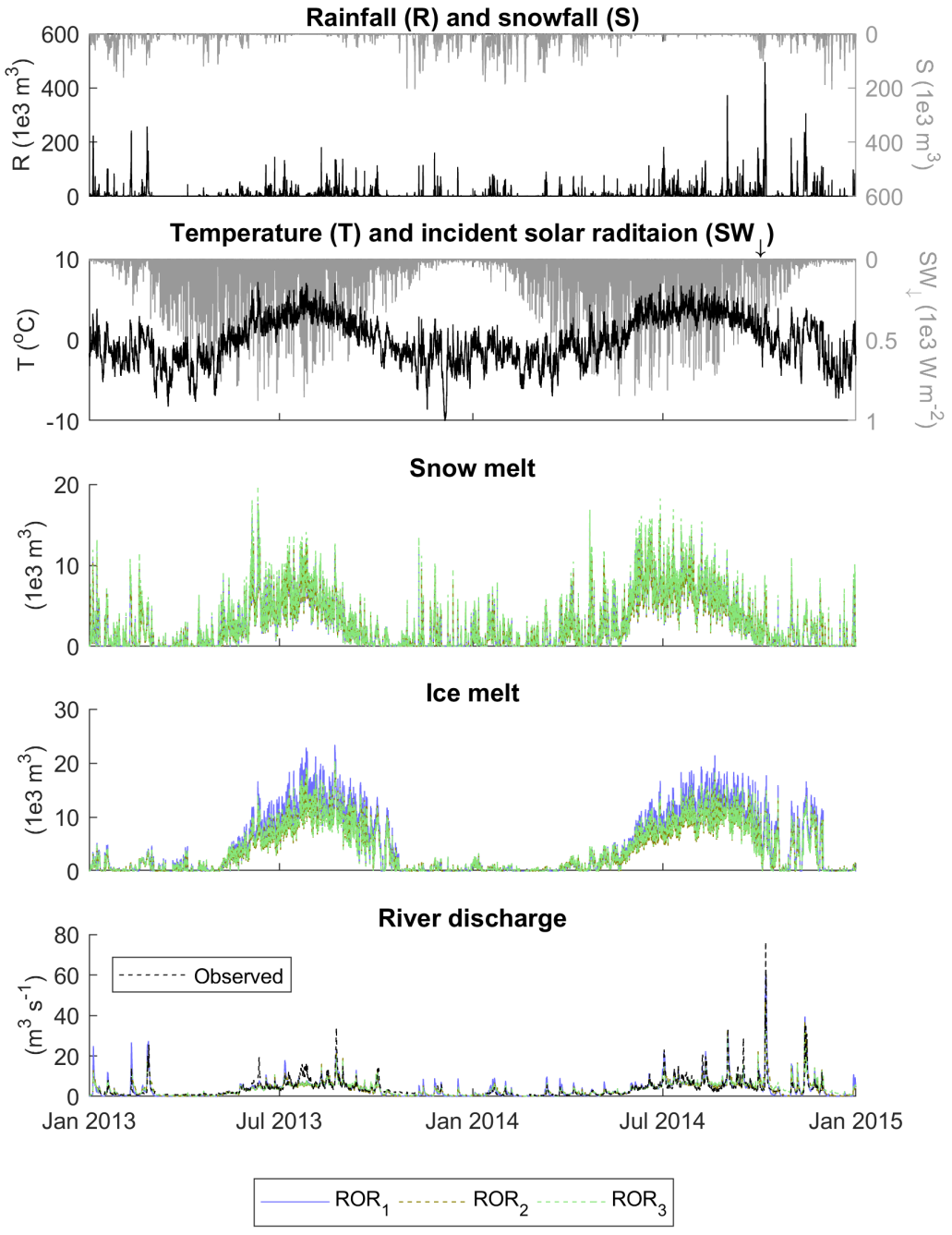

Figure D2. Time-series of driving precipitation, temperature and incident solar radiation data and simulated snow melt and ice melt and river discharge using the ROR_1, ROR_2 and ROR_3 runoff-routing model structures in conjunction with the simplest TIM_1 melt model structure. Note, the proportion of rainfall and snowfall is an output from the GHM which is approximately equal across the different configurations. Also note, ice melt includes melt of bare ice and the firn.

References

Aðalgeirsdóttir, G., Guðmundsson, S., Björnsson, H., Pálsson, F., Jóhannesson, T., Hannesdóttir, H., Sigurðsson, S. ., and Berthier, E.: Modelling the 20th and 21st century evolution of Hoffellsjökull glacier, SE-Vatnajökull, Iceland, The Cryosphere, 5, 961–975, https://doi.org/10.5194/tc-5-961-2011, 2011.

Allen, R., Pereira, L., Raes, D., and Smith, M.: Crop evapotranspiration - Guidelines for computing crop water requirements - FAO Irrigation and drainage paper 56, Tech. rep., Food and Agriculture Organization of the United Nations, Rome, Italy, 1998.

Andrés-Doménech, I., García-Bartual, R., Montanari, A., and Marco, J. B.: Climate and hydrological variability: The catchment filtering role, Hydrology and Earth System Sciences, 19, 379–387, https://doi.org/10.5194/hess-19-379-2015, 2015.

Arnold, N. S., Rees, W. G., Hodson, A. J., and Kohler, J.: Topographic controls on the surface energy balance of a high Arctic valley glacier, Journal of Geophysical Research: Earth Surface, 111, F02 011, https://doi.org/10.1029/2005JF000426, 2006.

Barnett, T. P., Adam, J. C., and Lettenmaier, D. P.: Potential impacts of a warming climate on water availability in snow-dominated regions., Nature, 438, 303–309, https://doi.org/10.1038/nature04141, 2005.

Barrand, N. E., Murray, T., James, T. D., Barr, S. L., and Mills, J. P.: Optimizing photogrammetric DEMs for glacier volume change assessment using laser-scanning derived ground-control points, Journal of Glaciology, 55, 106–116, https://doi.org/10.3189/002214309788609001, 2009.

Bengtsson, L., Andrae, U., Aspelien, T., Batrak, Y., Calvo, J., de Rooy, W., Gleeson, E., Hansen-Sass, B., Homleid, M., Hortal, M., Ivarsson, K.-I., Lenderink, G., Niemelä, S., Pagh Nielsen, K., Onvlee, J., Rontu, L., Samuelsson, P., Santos Muñoz, D., Subias, A., Tijm, S., Toll, V., Yang, X., and Ødegaard Køltzow, M.: The HARMONIE-AROME model configuration in the ALADIN-HIRLAM NWP system, Monthly Weather Review, pp. MWR–D–16–0417.1, https://doi.org/10.1175/MWR-D-16-0417.1, 2017.

Bergström, S.: Development and test of the distributed HBV-96 hydrological model, Journal of Hydrology, 201, 272–288, https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

Beven, K.: A manifesto for the equifinality thesis, Journal of Hydrology, 320, 18–36, https://doi.org/10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K.: Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, Hydrological Sciences Journal, 61, 1652–1665, https://doi.org/10.1080/02626667.2015.1031761, 2016.

Blaney, H. F. and Morin, K. V.: Evaporation and consumptive use of water empirical formulas, Eos, Transactions American Geophysical Union, 23, 76–83, https://doi.org/10.1029/TR023i001p00076, 1942.

Blazkova, S. and Beven, K.: A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, Water Resources Research, 45, W00B16, https://doi.org/10.1029/2007WR006726, 2009.

Boscarello, L., Ravazzani, G., Rabuffetti, D., and Mancini, M.: Integrating glaciers raster-based modelling in large catchments hydrological balance: The Rhone case study, Hydrological Processes, 28, 496–508, https://doi.org/10.1002/hyp.9588, 2014.

Bradwell, T., Sigurdsson, O., and Everest, J.: Recent, very rapid retreat of a temperate glacier in SE Iceland, Boreas, 42, 959–973, https://doi.org/10.1111/bor.12014, 2013.

Braithwaite, R. J.: Positive degree-day factors for ablation on the Greenland Ice-sheet studied by energy balance modeling, Journal of Glaciology, 41, 153–160, 1995.

Brately, P. and Fox, B. L.: Algorithm 659: Implementing Sobol's quasirandom sequence generator, ACM Transactions on Mathematical Software, 14, 88–100, 1988.

Brock, B. W., Willis, I. C., and Sharp, M. J.: Measurement and parameterisation of albedo variations at Haut Glacier d'Arolla , Switzerland, Journal of Glaciology, 46, 675–688, https://doi.org/10.3189/172756506781828746, 2000.

Cannon, A. J., Sobie, S. R., and Murdock, T. Q.: Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes?, Journal of Climate, 28, 6938–6959, https://doi.org/10.1175/JCLI-D-14-00754.1, 2015.

Carenzo, M., Pellicciotti, F., Mabillard, J., Reid, T., and Brock, B. W.: An enhanced temperature index model for debris-covered glaciers accounting for thickness effect, Advances in Water Resources, 94, 457–469, https://doi.org/10.1016/j.advwatres.2016.05.001, 2016.

Casper, M. C., Grigoryan, G., Gronz, O., Gutjahr, O., Heinemann, G., Ley, R., and Rock, A.: Analysis of projected hydrological behavior of catchments based on signature indices, Hydrology and Earth System Sciences, 16, 409–421, https://doi.org/10.5194/hess-16-409-2012, 2012.

Ciarapica, L. and Todini, E.: TOPKAPI: A model for the representation of the rainfall-runoff process at different scales, Hydrological Processes, 16, 207–229, https://doi.org/10.1002/hyp.342, 2002.

Clausen, B. and Biggs, B. J. F.: Flow variables for ecological studies in temperate streams: Groupings based on covariance, Journal of Hydrology, 237, 184–197, https://doi.org/10.1016/S0022-1694(00)00306-1, 2000.

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., and Clark, M.: Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments, Hydrological Processes, 28, 6135–6150, https://doi.org/10.1002/hyp.10096, 2014.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., and Smith, P. J.: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water Resources Research, 51, 5531–5546, https://doi.org/10.1002/2014WR016532, 2015.

de Woul, M., Hock, R., Braun, M., Thorsteinsson, T., Jóhannesson, T., and Halldórsdóttir, S.: Firn layer impact on glacial runoff: A case study at Hofsjökull, Iceland, Hydrological Processes, 20, 2171–2185, https://doi.org/10.1002/hyp.6201, 2006.

Duethmann, D., Menz, C., Jiang, T., and Vorogushyn, S.: Projections for headwater catchments of the Tarim River reveal glacier retreat and decreasing surface water availability but uncertainties are large, Environmental Research Letters, 11, 054 024, https://doi.org/10.1088/1748-9326/11/5/054024, 2016.

Einarsson, M. Á.: Evaporation and potential evapotranspiration in Iceland, Veðurstofa Íslands, Reykjavík, 1972.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G.: A framework to assess the realism of model structures using hydrological signatures, Hydrology and Earth System Sciences, 17, 1893–1912, https://doi.org/10.5194/hess-17-1893-2013, 2013.

Farinotti, D., Usselmann, S., Huss, M., Bauder, A., and Funk, M.: Runoff evolution in the Swiss Alps: projections for selected high-alpine catchments based on ENSEMBLES scenarios, Hydrological Processes, 26, 1909–1924, https://doi.org/10.1002/hyp.8276, 2012.

Finger, D., Pellicciotti, F., Konz, M., Rimkus, S., and Burlando, P.: The value of glacier mass balance, satellite snow cover images, and hourly discharge for improving the performance of a physically based distributed hydrological model, Water Resources Research, 47, W07 519, https://doi.org/10.1029/2010WR009824, 2011.

Finger, D., Vis, M., Huss, M., and Seibert, J.: The value of multiple data set calibration versus model complexity for improving the performance of hydrological models in mountain catchments, Water Resources Research, 51, 1939–1958, https://doi.org/10.1002/2014WR015712, 2015.

Flett, V. T.: Glacier retreat and projected river regime changes in the hydrologically highly-coupled Virkisjökull catchment , SE Iceland, Doctor of philosophy, University of Dundee, 2016.

Gabbi, J., Carenzo, M., Pellicciotti, F., Bauder, A., and Funk, M.: A comparison of empirical and physically based glacier surface melt models for long-term simulations of glacier response, Journal of Glaciology, 60, 1199–1207, https://doi.org/10.3189/2014JoG14J011, 2014.

Gao, H., Ding, Y., Zhao, Q., Hrachowitz, M., and Savenije, H. H.: The importance of aspect for modelling the hydrological response in a glacier catchment in Central Asia, Hydrological Processes, 31, 2842–2859, https://doi.org/10.1002/hyp.11224, 2017.

Garavaglia, F., Le Lay, M., Gottardi, F., Garçon, R., Gailhard, J., Paquet, E., and Mathevet, T.: Impact of model structure on flow simulation and hydrological realism: from lumped to semi-distributed approach, Hydrology and Earth System Sciences Discussions, pp. 1–21, https://doi.org/10.5194/hess-2017-82, 2017.

Gardner, A. S. and Sharp, M.: Sensitivity of net mass-balance estimates to near-surface temperature lapse rates when employing the degree-day method to estimate glacier melt, Annals of Glaciology, 50, 80–86, https://doi.org/10.3189/172756409787769663, 2009.

Gardner, A. S., Sharp, M. J., Koerner, R. M., Labine, C., Boon, S., Marshall, S. J., Burgess, D. O., and Lewis, D.: Near-surface temperature lapse rates over arctic glaciers and their implications for temperature downscaling, Journal of Climate, 22, 4281–4298, https://doi.org/10.1175/2009JCLI2845.1, 2009.

Griffiths, J., Keller, V., Morris, D., and Young, A.: Continuous Estimation of River Flows (CERF) : Model Scheme for Representing Rainfall Interception and Soil Moisture. Environment Agency R & D Project W6- 101., Tech. rep., Centre for Ecology and Hydrology, Wallingford, UK, 2006.

Guðmundsson, M. T.: Mass balance and precipitation on the summit plateau of Öræfajökull, SE-Iceland, Jökull, 48, 49–54, 2000.

Guðmundsson, S., Björnsson, H., Jóhannesson, T., Aðalgeirsdóttir, G., Pálsson, F., and Sigurðsson, O.: Similarities and differences in the response to climate warming of two ice caps in Iceland, Hydrology Research, 40, 495–502, https://doi.org/10.2166/nh.2009.210, 2009.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802–3813, https://doi.org/10.1002/hyp.6989, 2008.

Hannah, D. M. and Gurnell, A. M.: A conceptual, linear reservoir runoff model to investigate melt season changes in cirque glacier hydrology, Journal of Hydrology, 246, 123–141, https://doi.org/10.1016/S0022-1694(01)00364-X, 2001.

Hannesdóttir, H., Björnsson, H., Pálsson, F., Aðalgeirsdóttir, G., and Guðmundsson, S.: Changes in the southeast Vatnajökull ice cap, Iceland, between ˜ 1890 and 2010, The Cryosphere, 9, 565–585, https://doi.org/10.5194/tc-9-565-2015, 2015.

Hanzer, F., Helfricht, K., Marke, T., and Strasser, U.: Multilevel spatiotemporal validation of snow/ice mass balance and runoff modeling in glacierized catchments, Cryosphere, 10, 1859–1881, https://doi.org/10.5194/tc-10-1859-2016, 2016.

Heynen, M., Pellicciotti, F., and Carenzo, M.: Parameter sensitivity of a distributed enhanced temperature-index melt model, Annals of Glaciology, 54, 311–321, https://doi.org/10.3189/2013AoG63A537, 2013.

Hock, R.: A distributed temperature-index ice- and snowmelt model including potential direct solar radiation, Journal of Glaciology, 45, 101–111, 1999.

Hock, R. and Jansson, P.: Modeling Glacier Hydrology, in: Encyclopedia of Hydrological Sciences 4, edited by Anderson, M. G. and McDonnell, J. J., pp. 2647–2655, John Wiley & Sons, Ltd, Chichester, UK, 2005.

Hopkinson, C., Chasmer, L., Munro, S., and Demuth, M. N.: The influence of DEM resolution on simulated solar radiation-induced glacier melt, Hydrological Processes, 24, 775–788, https://doi.org/10.1002/hyp.7531, 2010.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H., and Gascuel-Odoux, C.: Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, Water Resources Research, 50, 7445–7469, 2014.

Huss, M., Bauder, A., Funk, M., and Hock, R.: Determination of the seasonal mass balance of four Alpine glaciers since 1865, Journal of Geophysical Research: Earth Surface, 113, F01 015, https://doi.org/10.1029/2007JF000803, 2008a.

Huss, M., Farinotti, D., Bauder, A., and Funk, M.: Modelling runoff from highly glacierized alpine drainage basins in a changing climate, Hydrological Processes, 22, 3888–3902, https://doi.org/10.1002/hyp.7055, 2008b.

Huss, M., Jouvet, G., Farinotti, D., and Bauder, A.: Future high-mountain hydrology: a new parameterization of glacier retreat, Hydrology and Earth System Sciences, 14, 815–829, https://doi.org/10.5194/hess-14-815-2010, 2010.

Huss, M., Zemp, M., Joerg, P. C., and Salzmann, N.: High uncertainty in 21st century runoff projections from glacierized basins, Journal of Hydrology, 510, 35–48, https://doi.org/10.1016/j.jhydrol.2013.12.017, 2014.

IGS: Icelandic Glaciological Society Terminus monitoring, http://spordakost.jorfi.is, 2017.

Immerzeel, W. W., Petersen, L., Ragettli, S., and Pellicciotti, F.: The importance of observed gradients of air temperature and precipitation for modeling runoff from a glacierized watershed in the Nepalese Himalayas, Water Resources Research, 50, 2212–2226, https://doi.org/10.1002/2013WR014506.Received, 2014.

IMO: Icelandic Meteorological Office and Institute of Earth Sciences, University of Iceland: DEMs of Icelandic glaciers (data set), 2013.

Irvine-Fynn, T. D. L., Hanna, E., Barrand, N. E., Porter, P. R., Kohler, J., and Hodson, A. J.: Examination of a physically based , high-resolution , distributed Arctic temperature-index melt model , on Midtre Lovénbreen , Svalbard, 28, 134–149, https://doi.org/10.1002/hyp.9526, 2014.

Jansson, P., Hock, R., and Schneider, T.: The concept of glacier storage: a review, Journal of Hydrology, 282, 116–129, https://doi.org/10.1016/S0022-1694(03)00258-0, http://linkinghub.elsevier.com/retrieve/pii/S0022169403002580, 2003.

Jeelani, G., Feddema, J. J., Van Der Veen, C. J., and Stearns, L.: Role of snow and glacier melt in controlling river hydrology in Liddar watershed (western Himalaya) under current and future climate, Water Resources Research, 48, https://doi.org/10.1029/2011WR011590, 2012.

Johannesson, T., Sigurdsson, O., Laumann, T., and Kennett, M.: Degree-day glacier mass-balance modelling with applications to glaciers in Iceland, Norway and Greenland, Journal of Glaciology, 41, 345–358, 1995.

Jost, G., Moore, R. D., Menounos, B., and Wheate, R.: Quantifying the contribution of glacier runoff to streamflow in the upper Columbia River Basin, Canada, Hydrology and Earth System Sciences, 16, 849–860, https://doi.org/10.5194/hess-16-849-2012, 2012.

Konya, K., Matsumoto, T., and Naruse, R.: Surface heat balance and spatially distributed ablation modelling at Koryto Glacier, Kamchatka peninsula, Russia, Geografiska Annaler, 86 A, 337–348, 2004.

Li, H., Sheffield, J., and Wood, E. F.: Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching, Journal of Geophysical Research Atmospheres, 115, D10 101, https://doi.org/10.1029/2009JD012882, 2010.

Li, H., Beldring, S., Xu, C.-Y., Huss, M., Melvold, K., and Jain, S. K.: Integrating a glacier retreat model into a hydrological model – Case studies of three glacierised catchments in Norway and Himalayan region, Journal of Hydrology, 527, 656–667, https://doi.org/10.1016/j.jhydrol.2015.05.017, 2015.

Lutz, a. F., Immerzeel, W. W., Shrestha, a. B., and Bierkens, M. F. P.: Consistent increase in High Asia's runoff due to increasing glacier melt and precipitation, Nature Climate Change, 4, 587–592, https://doi.org/10.1038/nclimate2237, 2014.

MacDougall, A. H., Wheler, B. A., and Flowers, G. E.: A preliminary assessment of glacier melt-model parameter sensitivity and transferability in a dry subarctic environment, Cryosphere, 5, 1011–1028, https://doi.org/10.5194/tc-5-1011-2011, 2011.

Mackay, J., Jackson, C., and Wang, L.: A lumped conceptual model to simulate groundwater level time-series, Environmental Modelling and Software, 61, 229–245, https://doi.org/10.1016/j.envsoft.2014.06.003, 2014.

Magnússon, E., Pálsson, F., Björnsson, H., and Guðmundsson, S.: Removing the ice cap of Öraefajokull central volcano, SE-Iceland: Mapping and interpretation of bedrock topography, ice volumes, subglacial troughs and implications for hazards assessments, Jökull, 62, 131–150, 2012.

Magnússon, E., Belart, J. M.-c., Pálsson, F., Ágústsson, H., and Crochet, P.: Geodetic mass balance record with rigorous uncertainty estimates deduced from aerial photographs and lidar data – Case study from Drangajökull ice cap , NW Iceland, The Cryosphere, 10, 159–177, https://doi.org/10.5194/tc-10-159-2016, 2016.

Matthews, T., Hodgkins, R., Wilby, R. L., Gumundsson, S., Pálsson, F., Björnsson, H., and Carr, S.: Conditioning temperature-index model parameters on synoptic weather types for glacier melt simulations, Hydrological Processes, 29, 1027–1045, https://doi.org/10.1002/hyp.10217, 2015.

Matthews, T. O. M. and Hodgkins, R.: Interdecadal variability of degree-day factors on Vestari Hagafellsjökull (Langjökull, Iceland) and the importance of threshold air temperatures, Journal of Glaciology, 62, 310–322, https://doi.org/10.1017/jog.2016.21, 2016.

Mayr, E., Hagg, W., Mayer, C., and Braun, L.: Calibrating a spatially distributed conceptual hydrological model using runoff, annual mass balance and winter mass balance, Journal of Hydrology, 478, 40–49, https://doi.org/10.1016/j.jhydrol.2012.11.035, 2013.

McMillan, H. K. and Westerberg, I. K.: Rating curve estimation under epistemic uncertainty, Hydrological Processes, 29, 1873–1882, https://doi.org/10.1002/hyp.10419, 2015.

Minder, J. R., Mote, P. W., and Lundquist, J. D.: Surface temperature lapse rates over complex terrain: Lessons from the Cascade Mountains, Journal of Geophysical Research Atmospheres, 115, D14 122, https://doi.org/10.1029/2009JD013493, 2010.

Monk, W. A., Wood, P. J., Hannah, D. M., and Wilson, D. A.: Selection of river flow indices for the assessment of hydroecological change, River Research and Applications, 23, 113–122, https://doi.org/10.1002/rra.964, 2007.

Mosier, T. M., Hill, D. F., and Sharp, K. V.: How much cryosphere model complexity is just right? Exploration using the conceptual cryosphere hydrology framework, Cryosphere, 10, 2147–2171, https://doi.org/10.5194/tc-10-2147-2016, 2016.

Nawri, N., Pálmason, B., Petersen, G. N., Björnsson, H., and Þorsteinsson, S.: The ICRA atmospheric reanalysis project for Iceland, Tech. rep., Icelandic Meteorological Office, Reykjavík, Iceland, 2017.

Nepal, S., Flügel, W.-A., Krause, P., Fink, M., and Fischer, C.: Assessment of Spatial Transferability of Process-Based Hydrological Model Parameters in Two Neighboring Catchments in the Himalayan Region, Hydrological Processes, pp. 1–15, https://doi.org/10.1002/hyp.11199, 2017.

Oerlemans, J.: Glaciers and Climate Change, A. A. Balkema Publishers, Rotterdam, Netherlands, 2001.

Ohmura, A.: Physical Basis for the Temperature-Based Melt-Index Method, Journal of Applied Meteorology, 40, 753–761, https://doi.org/10.1175/1520-0450(2001)040<0753:PBFTTB>2.0.CO;2, 2001.

Pappenberger, F., Matgen, P., Beven, K. J., Henry, J. B., Pfister, L., and Fraipont, P.: Influence of uncertain boundary conditions and model structure on flood inundation predictions, Advances in Water Resources, 29, 1430–1449, https://doi.org/10.1016/j.advwatres.2005.11.012, 2006.

Pellicciotti, F., Brock, B., Strasser, U., Burlando, P., Funk, M., and Corripio, J.: An enhanced temperature-index glacier melt model including the shortwave radiation balance : development and testing for Haut Glacier d ' Arolla , Switzerland, 51, 573–587, 2005.

Pellicciotti, F., Helbing, J., Rivera, A., Favier, V., Corripio, J., Araos, J., Sicart, J.-E., and Carenzo, M.: A study of the energy balance and melt regime on Juncal Norte Glacier, semi-arid Andes of central Chile, using melt models of different complexity, Hydrological Processes, 22, 3980–3997, 2008.

Pellicciotti, F., Buergi, C., Immerzeel, W. W., Konz, M., and Shrestha, A. B.: Challenges and Uncertainties in Hydrological Modeling of Remote Hindu Kush–Karakoram–Himalayan (HKH) Basins: Suggestions for Calibration Strategies, Mountain Research and Development, 32, 39–50, https://doi.org/10.1659/MRD-JOURNAL-D-11-00092.1, 2012.

Petersen, L. and Pellicciotti, F.: Spatial and temporal variability of air temperature on a melting glacier: Atmospheric controls, extrapolation methods and their effect on melt modeling, Juncal Norte Glacier, Chile, Journal of Geophysical Research Atmospheres, 116, D23 109, https://doi.org/10.1029/2011JD015842, 2011.

Phillips, E., Finlayson, A., Bradwell, T., Everest, J., and Jones, L.: Structural evolution triggers a dynamic reduction in active glacier length during rapid retreat: Evidence from Falljökull, SE Iceland, Journal of Geophysical Research F: Earth Surface, 119, 2194–2208, https://doi.org/10.1002/2014JF003165, 2014.

Ponce, V. M.: Engineering hydrology: Principles and practices, http://ponce.sdsu.edu/enghydro/, 2014.

Radić, V. and Hock, R.: Glaciers in the Earth's Hydrological Cycle: Assessments of Glacier Mass and Runoff Changes on Global and Regional Scales, Surveys in Geophysics, 35, 813–837, https://doi.org/10.1007/s10712-013-9262-y, 2014.

Ragettli, S., Cortés, G., Mcphee, J., and Pellicciotti, F.: An evaluation of approaches for modelling hydrological processes in high-elevation, glacierized Andean watersheds, Hydrological Processes, 28, 5674–5695, https://doi.org/10.1002/hyp.10055, 2014.

Ragettli, S., Immerzeel, W. W., and Pellicciotti, F.: Contrasting climate change impact on river flows from high-altitude catchments in the Himalayan and Andes Mountains., Proceedings of the National Academy of Sciences of the United States of America, 113, 9222–9227, https://doi.org/10.1073/pnas.1606526113, 2016.

Reda, I. and Andreas, A.: Solar Position Algorithm for Solar Radiation Applications, Tech. Rep. NREL/TP-560-34302, National Renewable Energy Laboratory, Colorado, USA, 2008.

Reveillet, M., Vincent, C., Six, D., and Rabatel, A.: Which empirical model is best suited to simulate glacier mass balances?, Journal of Glaciology, 63, 39–54, https://doi.org/10.1017/jog.2016.110, 2017.

Riggs, G. and Hall, D.: MODIS Snow Products Collection 6 User Guide, Tech. rep., 2015.

Rye, C. J., Willis, I. C., Arnold, N. S., and Kohler, J.: On the need for automated multiobjective optimization and uncertainty estimation of glacier mass balance models, Journal of Geophysical Research: Earth Surface, 117, 1–21, https://doi.org/10.1029/2011JF002184, 2012.

Sachindra, D. A., Huang, F., Barton, A., and Perera, B. J. C.: Statistical downscaling of general circulation model outputs to precipitation-part 2: Bias-correction and future projections, International Journal of Climatology, 34, 3282–3303, https://doi.org/10.1002/joc.3915, 2014.

Salomonson, V. V. and Appel, I.: Estimating fractional snow cover from MODIS using the normalized difference snow index, Remote Sensing of Environment, 89, 351–360, https://doi.org/10.1016/j.rse.2003.10.016, 2004.

Sawicz, K. A., Kelleher, C., Wagener, T., Troch, P., Sivapalan, M., and Carrillo, G.: Characterizing hydrologic change through catchment classification, Hydrology and Earth System Sciences, 18, 273–285, https://doi.org/10.5194/hess-18-273-2014, 2014.

Schaefli, B.: Snow hydrology signatures for model identification within a limits-of-acceptability approach, Hydrological Processes, 30, 4019–4035, https://doi.org/10.1002/hyp.10972, 2016.

Schaefli, B., Nicótina, L., Imfeld, C., Da Ronco, P., Bertuzzo, E., and Rinaldo, A.: SEHR-ECHO v1.0: A spatially explicit hydrologic response model for ecohydrologic applications, Geoscientific Model Development, 7, 2733–2746, https://doi.org/10.5194/gmd-7-2733-2014, 2014.

Schulla, J.: Model Description WaSiM. Technical report., Tech. rep., Hydrology Software Consulting, Zürich, 2015.

Shafii, M. and Tolson, B. A.: Optimizing hydrological consistency by incorporating hydrological signatures intomodel calibration objectives, Water Resources Research, 51, 3796–3814, 2015.

Shea, J. M. and Moore, R. D.: Prediction of spatially distributed regional-scale fields of air temperature and vapor pressure over mountain glaciers, Journal of Geophysical Research Atmospheres, 115, D23 107, https://doi.org/10.1029/2010JD014351, 2010.

Singh, S., Kumar, R., Bhardwaj, A., Sam, L., Shekhar, M., Singh, A., Kumar, R., and Gupta, A.: Changing climate and glacio-hydrology in Indian Himalayan Region: A review, Wiley Interdisciplinary Reviews: Climate Change, 7, 393–410, https://doi.org/10.1002/wcc.393, 2016.

Sorensen, J. P. R., Finch, J. W., Ireson, A. M., and Jackson, C. R.: Comparison of varied complexity models simulating recharge at the field scale, Hydrological Processes, 28, 2091–2102, https://doi.org/10.1002/hyp.9752, 2014.

Srivastav, R. K., Schardong, A., and Simonovic, S. P.: Equidistance Quantile Matching Method for Updating IDF Curves under Climate Change, Water Resources Management, 28, 2539–2562, https://doi.org/10.1007/s11269-014-0626-y, 2014.

Switanek, M. B., Troch, P. A., Castro, C. L., Leuprecht, A., Chang, H.-I., Mukherjee, R., and Demaria, E. M. C.: Scaled distribution mapping: a bias correction method that preserves raw climate model projected changes, Hydrology and Earth System Sciences, 21, 2649–2017, 2017.

Teutschbein, C., Grabs, T., Karlsen, R. H., Laudon, H., and Bishop, K.: Hydrological response to changing climate conditions: Spatial streamflow variability in the boreal region, Water Resources Research, 51, 9425–9446, https://doi.org/10.1002/2015WR017337, 2015.

Van Tiel, M., Teuling, A. J., Wanders, N., Vis, M. J. P., Stahl, K., and Van Loon, A. F.: The role of glacier dynamics and threshold definition in the characterisation of future streamflow droughts in glacierised catchments, Hydrology and Earth System Sciences Discussions, pp. 1–31, https://doi.org/10.5194/hess-2017-119, 2017.

Viglione, A., Parajka, J., Rogger, M., Salinas, J. L., Laaha, G., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins - Part 3: Runoff signatures in Austria, Hydrology and Earth System Sciences, 17, 2263–2279, https://doi.org/10.5194/hess-17-2263-2013, 2013.

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., and Freer, J.: Uncertainty in hydrological signatures for gauged and ungauged catchments, Water Resources Research, 52, 1847–1865, https://doi.org/10.1002/2015WR017635, 2016.

Winsemius, H. C., Schaefli, B., Montanari, A., and Savenije, H. H. G.: On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information, Water Resources Research, 45, W12 422, https://doi.org/10.1029/2009WR007706, 2009.

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Advances in Water Resources, 30, 1756–1774, https://doi.org/10.1016/j.advwatres.2007.01.005, 2007.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44, W09 417, https://doi.org/10.1029/2007WR006716, 2008.

Zhang, Y., Hirabayashi, Y., Liu, Q., and Liu, S.: Glacier runoff and its impact in a highly glacierized catchment in the southeastern Tibetan Plateau: past and future trends, Journal of Glaciology, 61, 713–730, https://doi.org/10.3189/2015JoG14J188, 2015.