Below, find the original reviewer concern/question in regular typeface followed by our response in italic typeface. This response is written in anticipation of making all corrections to a manuscript as stated. Therefore, we make statements such as 'we changed' or 'we corrected' or 'we modified' in past-tense as if this document were being submitted together with a revised manuscript. It is our intent that the revised manuscript will read exactly as suggested here. We have also included references to the other review.

Reviewer:
This paper details a study using the Hillslope River Routing Model (HRR) to better constrain surface meltwater routing on Greenland, improving on results that only use surface mass balance models. This is a good example of taking more advance techniques used in hydrology and applying them to glaciology, something which can greatly benefit the field.

However, I don't find the way in which the paper is currently written to be suitable for the target audience of a journal such as The Cryosphere. Many of the terms and methods which may be more familiar to trained hydrologists will not be familiar to many glaciologists. I would very much like to see this paper published in The Cryosphere as it would be a good step towards better collaboration between these fields, but I would recommend a detailed re-write to make it more accessible to those without a background in hydrology.

We thank the reviewer for this perspective, and indeed our collaboration of a hydrologist (lead author) and Greenland specialists (most other authors) resulted in many rounds of internal editing and terminology clarification. We targeted TC for this submission because we felt it was most novel and interesting to the glaciology community, and we are glad the reviewer agrees with that decision.

We have made many changes to the text to better define jargon and standard practices from hydrology for the glaciology community, and the reviewer's detailed comments where helpful in guiding where these passages might be needed. When a full and satisfying explanation might have taken too much text In making these changes, we have referred readers to classic citations (while still including hopefully enough information here so the reader doesn't have to look it up). Our changes are highlighted as responses to the specific concerns below.

My detailed comments are below:
Line 34: and lakes?
We sought to define and distinguish the emerging rivers and streams research against the more established study of lakes here, although we agree that the study of supraglacial hydrology as a whole is an emerging discipline. The paragraph beginning on line 61 covers relevant lake research, including some new papers suggested by both reviewers (see below).

Line 40: Also Leeson et al. 2012 DOI: 10.5194/tc-6-1077-201
Cited as suggested by both reviewers. Thanks for noting this omission.

Line 82: What about subglacial channels/ transport through firn?
We have included the parenthetical "(which include firn atop the GrIS)" in this line to establish that 'hillslope' soil processes are akin to firn/bare ice process in Greenland.

Line 97: Is this the only data available for the whole of Greenland? It seems a big weakness of this paper is the lack of validation data available. If other datasets were not used it should be explained why (but if this is really all there is then fair enough!).

This is, to our knowledge, the only validation available in the whole of Greenland. These data are very difficult to obtain in the field, and while there are a few other observations of river discharge atop the GrIS (discussed in lines 122-125), the Smith et al data are the only available source sufficient for calibration of a hydraulic model as performed here. We agree that this is a weakness of this paper, but hope that this paper spurs other authors (and funders) to collect more in situ flow records to push this research forward.

Line 122: Related to the previous comment, is there data for other times available? Can you explain why it is so important only the peak of the melt-season can be used?
We see that our writing was confusing! It is not necessary that peak melt season data are used, and the methods should work at any time. Per previous work, we expect peak melt season to correspond to the maximum drainage network extent and therefore the most efficient drainage of the ice sheet (and thus most likely to agree with SMB modelling). We have eliminated this phrase to avoid similar confusion.

Line 129-133: Were other images available? What if other streams are present and just didn't have water flowing on this particular day?
The stream network from Smith et al (2017) was investigated by Yang et al (2018) as cited, and that work showed that the peak melt season image we re-use here produced the most extensive stream network and that channels were visible in the image. Since we couple this image to the DEM, and the DEM is ultimately the product that determines how many streams are present, we are not reliant on the representativeness of this single image as the DEM should generate channels not captured in the image. That said, and toward a question the reviewer asks later, the narrowest streams here are ~0.2m, and thus this 0.5m imagery is not missing any features that would have a major effect on flow routing.

This passage has been edited as follows: "Smith et al.'s stream network product was combined with a seasonally simultaneous portion of the 2 m resolution ArcticDEM DEM (Digital Elevation Model) obtained from Polar Geospatial Centre (Porter et al., 2018) to produce two distinct supraglacial stream networks as described in Section 3.2."

Line 134: Please define DEM
Defined as requested

Line 142: How do you determine all of the water in each of these grid cells will remain in this catchment, could any of it be being transported elsewhere?
This is an excellent question. Terrestrial hydrology theory holds that all water at the surface or in the near surface (in the GrIS- in a river, lake, or the firn/crust) is topographically controlled. That is, the watershed as defined by elevation by definition contains all transport as water flows 'down-gradient'. This assumption could be incorrect in several cases that are not present in terrestrial hydrology: 1) if water penetrates the firn into an englacial system driven by pressure heads and not topography, 2) if the channel network in reality thermally erodes some topography to the point where it 'breaches' the watershed divide, or 3) if there are large scale subsurface fluxes that transport subsurface water into or out of the system. In terrestrial hydrology, these subsurface fluxes do exist, leading to discussions on the 'myth of the water balance' in the literature. However, for scales such as the catchment here these fluxes are extremely uncommon. In the absence of literature describing englacial fluxes at medium to GrIS

scales (which in fairness would be extraordinarily difficult to measure), we are confident in this classical assumption and believe the 3rd exception is not valid. Further, the remotely sensed image of streams shows the stream network as it actually exists (as opposed to the traditional hydrologic view driven solely by a DEM that would assume topography), and therefore we can eliminate the 2nd exception. We are left with the 1st exception as a plausible means to violate our assumption of topographically controlled transport.

To state this assumption formally, we add the following to this passage: "which we assume is topographically constrained and transported exclusively via surface/near-surface transport" . We also state (per other suggestions by both reviewers) the following: "Further, the use of a single R_{coef} *allows us to accurately model discharge without allowing attribution of errors in runoff production: these could stem from SMB errors, unaccounted for refreezing, storage, or lake filling, surface transport that violates topographic constraints, englacial draining, or ADCP measurement error. Our framework is unable to apportion any gaps in runoff production and routed discharge to any of these sources, and thus our treatment of runoff as a bulk reduction/augmentation is faithful to our experiment design and manuscript goals."*

Line 167: Please define Froude number.
The Froude number is a classic hydraulic ratio of velocity and celerity in an open channel. The Froude number determines the flow regime (sub critical, critical, or supercritical), and these regimes are used in canal and river restoration design. Froude numbers essentially indicate whether surface waves propagate upstream or downstream given the balance of gravity driven flow and energy driven flow from upstream.

Per the reviewer's suggestions to improve the use of hydrologic jargon, we have added "(a classic index of flow velocity in open channel hydraulics)" after the introduction of the term.

Line 163 onwards: Why was none of this data suitable to use to validate the model?
Great question. In order to uses these data to validate the model, we would have to set it up in a controlled experiment: we would put the correct runoff forcing upstream of only these measurements and then route just the appropriate subbasin to produce the model estimates of these hydraulics at the appropriate location. We could have used these measurements if they were made at the same location at the ADCP. The internal model calculations that produce the hydraulics must by definition match these measurements at a point given conservation of mass if the input is correct, so this poses a problem given the lack of process modelling and mass gaps we discuss extensively later.

In further service of jargon translation, we now close this paragraph with "Note that we cannot use these observations to validate our routing model, and instead use them to inform it. These point measurements could in theory be reproduced by our hydraulic model, but to do so would require measurements of channel properties and runoff upstream of each point for several hours/days before each hydraulic measurement was taken, and such data do not exist."

Line 176: I know what you're getting at here but I'm not sure 'factorial' is the correct description?
Since this has sown confusion, we have dropped this and instead state the simpler ("four runs per model") in this passage.

Line 185: Is it possible that the SMB values are correct and water just isn't making it to the channels where the ADCP measurements are made? E.g. What about refreezing in firn?

This is possible, but we cannot say for sure, and it would seem a large amount of mass to refreeze. It could also be that a slowly draining englacial system is in play, or refreezing in the bottom of a lake, but we have seen no evidence of these factors in this watershed. It is true that we observe less outflow than SMB runoff would predict. Since we do not know why this might be the case, the use of R_{coef} *allows us to account for this difference without attribution. We do not want to make any conclusions we cannot support, and our modelling setup does not let us state anything beyond what we observed- there is a mismatch between how much mass is produced and how much is conserved to pass the outlet with correct timing. Reviewer 2 had similar concerns (although their focus on was lake filling and englacial drainage), so we have substantially rewritten the discussion and methods. We copy the new text below*

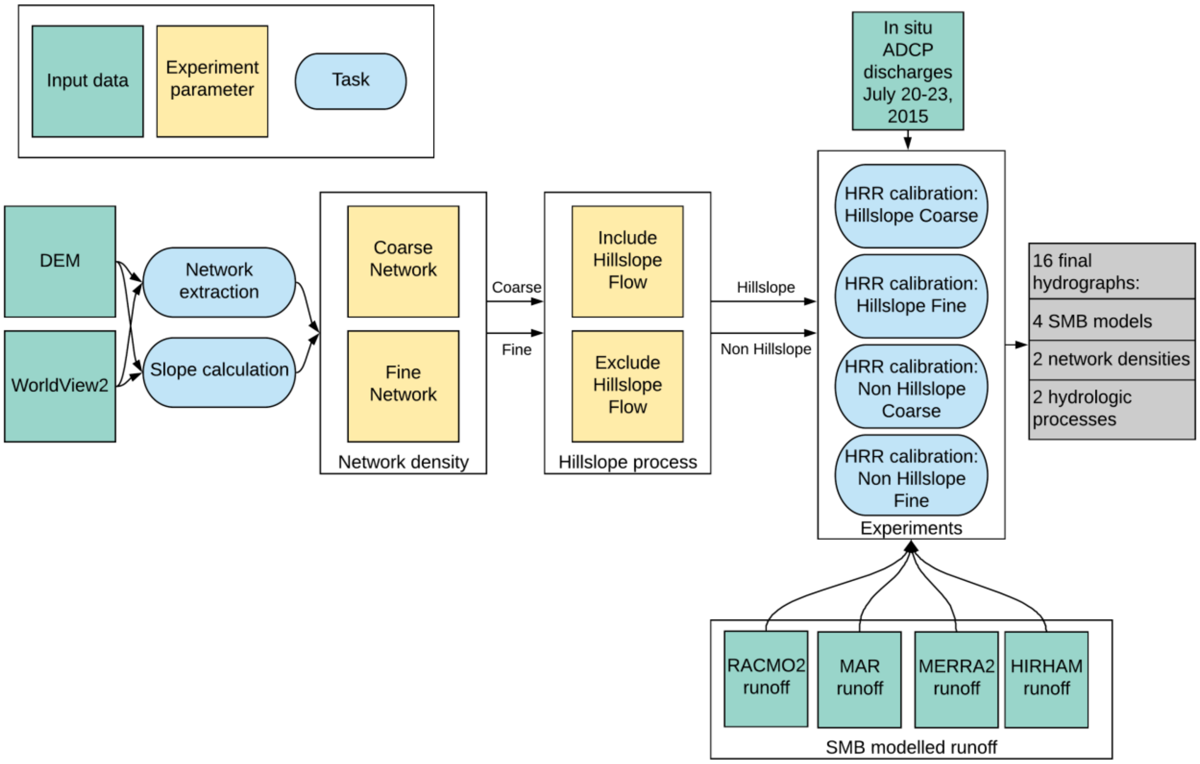
We now state in 3.1: "Further, the use of a single R_{coef} *allows us to accurately model discharge without allowing attribution of errors in runoff production: these could stem from SMB errors, unaccounted for refreezing, storage, or lake filling, surface transport that violates topographic constraints, englacial draining, or ADCP measurement error. Our framework is unable to apportion any gaps in runoff production and routed discharge to any of these sources, and thus our treatment of runoff as a bulk reduction/augmentation is faithful to our experiment design and manuscript goals."*

And in the discussion [entire paragraph reprinted]:

"Our results also support earlier assertions of mismatched timing and magnitude of SMB runoff and observed discharges entering the Rio Behar terminal moulin (Smith et al., 2017). The routing model is unable to assign glaciologic process to mass gaps, so we can only suggest plausible mechanisms for closing that mass balance gap. Mass gaps could perhaps result from to subsurface retention and/or refreezing in bare-ice weathering crust (Cooper et al., 2018), a process not currently well-represented in SMB models. Or, the mass imbalance could come from transport processes: filling lakes, drainage through fractures (there are no crevasses in the study area), or the breach of topographic divides are all plausible transport process gaps. Topographic breach is unlikely given that we use an observed (via image) channel network, and thus if breaches did occur they are accounted for. Further, total depression storage (including true lakes and DEM artifacts) was $6.92e^{6}$ m^{3}*, which is two orders of magnitude less than the observed ADCP flux during this time (integrated into a bulk volume,* $241e^{6}$ m^{3}*) and one order of magnitude less than the maximum runoff deficit (obtained by subtracting the ADCP from the largest SMB input,* $17.5\ e^{6}$ m^{3}*). Therefore, if all depressions were dry at the start of routing and were completely filled by runoff before beginning to flow in the channel network, this would still only account for roughly one-third of extra runoff production mass. Given that we know lakes are full during this time period, we assert that this lake filling effect is not the cause of mass imbalance. Further, errors in our outlet hydrographs are dominated by underestimation of night-time low flow periods, as peak flows are modelled well across nearly all 16 trials. These night-time low flows are particularly important for mass balance in the Rio Behar watershed, as a large driver of mismatches in total mass balance (Figure 4) comes from these low flow periods. Error could come from the ADCP itself, and this instrument is generally less certain at lower flows. However, the ADCP record here is taken from Smith et al. (2017) and represents a well documented procedure carried out by expert field personnel, and thus we are confident that ADCP errors are too small to explain* R_{coef}*. We affirm that all SMB models examined here produce too much excess water relative to ADCP observations (at least at peak times, Figure 4 shows MERRA2 total runoff is less than the ADCP total discharge, but still requires* $R_{coef} < 1$ *to reduce the peak daytime volume of water), and do not model night-time flows without routing, consistent with Smith et al. (2017). Our results suggest that hydrologic process modelling (i.e., routing) can correctly reproduce these night-time low flows."*

Figure 1: The middle 4 squares could be laid out in a clearer way. Why does including or excluding hillslope flow have a 'yes' or 'no' but there is no differentiation between coarse and fine network?

Good question. We have streamlined the figure as follows, which hopefully communicates more clearly.



Line 195-6: 'generalizing the process of water routing from satellite image collection to water routing' doesn't make sense to me.

Excellent point, as this was written in a way that doesn't make much sense upon rereading, and Reviewer 2 was also confused. We meant that we wanted to build a general and flexible process that can start with a remotely sensed channel map and take that map all the way to hourly routing (given calibration data). This passage is now stated "Although Smith et al. (2017) provide a topologically connected channel network for our study area (i.e., they explicitly defined how every channel is connected to every other channel throughout the entire network to allow water to flow from the headwaters to the outlet to obey observed channel connections) , we are interested in generalizing the process of water routing from satellite image collection to water routing in cases where pre-existing channel network maps do not exist., Further, we must generate different river networks to which is also necessary to test the effects of network density on the routing model."

Line 195 onwards: This paragraph was one I especially felt could be better explained for the non-hydrologist. It would also be good to see justification for 'standard hydrologic' practices being applied to icy surfaces.

Understood! We have added hopefully clarifying passages that have made this paragraph quite a bit longer (we split it into multiple paragraphs), but hopefully more legible. Since this was a major point of confusion, we have drafted revised text and reprint it here:

“Although Smith et al. (2017) provide a topologically connected channel network for our study area (i.e., they explicitly defined how every channel is connected to every other channel throughout the entire network to allow water to flow from the headwaters to the outlet to obey observed channel connections) , we are interested in generalizing the process of water routing in cases where pre-existing channel network maps do not exist. Further, we must generate different river networks to test the effects of network density on the routing model. Therefore, we introduce a process to create models of a complete river networks as defined by topography that can in theory be applied to any area of the GrIS with a high-quality DEM and a remotely sensed image. This topographically defined flow is a classic practice in terrestrial hydrology, and since all open channel flow is gravity-driven this practice applies for flow routing through any media without substantial pressure forces. Topographically defined flow has therefore been applied/invoked for a variety of surfaces, including Mars (e.g. Dohm et al., 2001; Rodriguez et al., 2005; Fassett and Head, 2008)

To generate our river networks, 1) We first ‘burned’ (i.e. lowered the pixel elevations) the remotely sensed stream map of Smith et al. (2017) into ArcticDEM, a standard hydrologic practice (e.g. Lindsay, 2016). This process ensures that channels are lower than surrounding topography, as remotely sensed DEMs cannot ‘see’ channel bottoms and therefore creates smooth surfaces where surface water features exist. Since we know that a river/stream channel is abruptly deeper than its surrounding banks, artificially lowering elevations where we observe channels ensures that these locations are the lowest feature in the surrounding terrain and therefore collect topographically driven water. In DEM processing for a hydrology, a depression is an area where water pools, as the flow direction is always downhill as in the sides of a bowl. These depressions typically need to be artificially ‘filled’, that is, their elevations need to be raised, as otherwise the topography indicates that water cannot leave once it enters the depression. Because we ‘burn in’ stream locations to the DEM, standard sink filling is not required for this analysis (we lower streams rather than raise depressions), but two large topographic depressions in the DEM of our catchment required further processing even after burning in streams. Standard DEM preparation for network generation dictates that upstream depressions are filled while an outlet depressions are preserved, yet this assumption generated unrealistic parallel drainage channels upstream and no channels in the outlet depression for our data. To address this problem, 2) a priority-flood algorithm (Lindsay, 2016) was applied to breach the two depressions and to create a continuously-flowing, realistic drainage network for the catchment (Figure 2). Finally, 3) the parameter that drives network generation and ultimately channel density is the channel initiation threshold: the minimum area needed to form a free-flowing channel. This concept stems from the fact that above river headwaters, water simply flows through the soil and not on the surface until the water table elevation exceeds the soil elevation in a spring. We observe an exact analogue on the GrIS: channels dwindle in size until they become indistinguishable from wet firn/ice near topographic divides (Gleason et al., 2016). To estimate the impact of drainage pattern on meltwater routing, we tested both a large (10^4 m^2) and a small (10^3 m^2) channel initiation threshold to create a ‘coarse’ and a ‘fine’ supraglacial drainage network, respectively from the DEM (Figure 2). These two modelled stream networks both follow the channel map from Smith et al (2017), with the key difference that the coarse network does not produce the narrowest streams we know to exist. This enabled us to test the effects of including or excluding very small tributary streams on surface water routing. We assign channel widths to each DEM-derived channel from the Smith et al channel map, and since the DEM process begins with burning in these streams, there is always a 1:1 assignment of channel width from imagery to network model. Our ‘fine’ channel network produces streams with a minimum width of 0.5m, matching to the correct order magnitude Gleason et al.’s (2016) reporting of channels as narrow as 0.2m. The coarse network produced streams with a minimum width of 0.7m, suggesting it is excluding the smallest streams in the remotely sensed map. GrIS supraglacial channels incise and meander over time, yet HRR cannot represent this behaviour and *instead assumes that channels remain fixed in space and time. It would be possible to derive expected erosion and incision (and additional meltwater) due to frictional heating of the channels, but without including a radiation budget and ice property data we could not model how the stream network changes in time nor satisfactorily model this additional meltwater with commensurate sophistication to the SMB runoff forcing (i.e., tight coupling with SMB models). Instead, we model these network snapshots with HRR as loosely coupled to SMB runoff (as opposed to tightly coupled, where SMB runoff would be an input into network generation), which is reasonable for our one-month experiment (Section 3.3.1)."*

Line 210-11: 0.5 is more than twice 0.2 so I wouldn't say they are matching, does this difference matter?
Thanks for pointing it out- what is important is the order of magnitude matching, as the differences between a stream of 0.2m and 0.5m are negligible for routing. Also, since remotely sensed stream widths are directly appended to the DEM generated channels (see updated text quoted above), this means simply that the process is capturing streams down to the limit of the imagery, which is a positive result.

Line 217 onwards: Again this paragraph could be clearer. The authors discuss stream order again in the results so an introduction to what this means and why this matters would be helpful.
Great suggestion. We have now included citation to a classic reference with explanation : "The coarse network has six stream orders (e.g. the smallest streams on the landscape are defined as order 1, and every junction of stream produces a new stream of higher order), and the fine network seven orders. Stream orders are a shorthand for hydraulic complexity of a network as the number and length of streams in a given order both increase geometrically (Horton, 1945). Therefore, our finding of almost an order of magnitude more channels in the fine 7-order network than the course 6-order network matches theory"

Line 234: Do any other models have this rigorous representation? If they are written in other, more accessible, languages it may still be useful to briefly mention them for those who may want to build on this but not use FORTRAN.
There are essentially only two options for efficient flow routing 'off the shelf' at large scales like these in current terrestrial hydrology: HRR, and RAPIDE (David et al., 2011). RAPIDE uses Muskingum routing instead of Muskingum-Cunge routing and does not contain a hillslope model, which makes it inferior for this application as we can't simulate hillslope (e.g. firn and bare ice) flow separately per subcatchment. In our view, the use of FORTRAN is a strength- it is an open source language that computes very quickly. If we were to recode HRR in Python, R, or Julia (considering other common open source languages), it would run orders of magnitude more slowly. We wrap HRR in an R framework- we invoke its FORTRAN core from RStudio, which allows for an easily shared workflow.

Line 245: Please give a little more detail about how EAs work.
We want include the intent of the reviewer without diving too deeply into the very large literature on hydrologic model calibration, so we have included the following passage "A very large literature on hydrologic model optimization and calibration exists, and interested readers are referred to Kirshner (2006) and Gupta et al (1998) for broad overviews of the subject. We do perform calibration this using an established evolutionary algorithm (EA; NSGA II, Deb et al., 2002) as EAs are efficient estimators in large parameter spaces that can achieve near-optimal results (Gleason and Smith, 2014)." The citations we give here are classic point and counterpoint papers that approach the subject of model calibration in a high level discourse. The Gupta school speaks to the art of calibration and the ability to improve

outcomes, while the Kirshner school bemoans that model calibration has become an art unto itself that obscures true understanding.

Line 255: Please define Manning's n.
We have added the following text in this passage:
" Channel friction is represented by Manning's n *and the EA solves for a single n per bin and assigns that n to all streams falling within that drainage area threshold. Manning (1891) generalized open channel flow into a simple equation where all flow resistances are lumped into a single empirical parameter n, and over a century of subsequent research has related n to landscape variables, channel form, and other geomorphic controls. Our binning of Manning's* n *follows general hydraulic correlations between channel size, slope, total discharge, and* n *(Brinkerhoff et al., 2019)."*

Line 296-7: Please define acronyms.
The hydrological acronym has been previously defined (NSE in line 275).

Line 311 and figure 4: Please comment more on why we are assuming here that MERRA2 is incorrect. Could it not be that all the other inputs are less accurate and MERRA2 is actually getting it right?
We didn't assume any models were correct/incorrect in this study in the beginning, but previous literature suggested we might need to adjust runoff. We took the published SMB output and compared it to field measurements before and after routing, and used the parameters we solved for (especially R_{coef}*) as a guide for how much correction each model requires to match observations. We do assume that our ADCP is correct, but this doesn't mean that the SMB model is incorrect: there is a process gap between what the models produce and what we observe, and this gap cannot be closed by flow routing. That gap is highlighted for the first time in this manuscript. MERRA2 has the same timing errors as all the other models, although it has a much better order of magnitude of flow. However, while MERRA2 is the closest 'out of the box' it is not the most accurate model after routing, and ultimately none of the models are correct in matching the ADCP without routing.*

Line 402 onwards: How is the slow lateral transport accounted for in the model?
Lateral transport is accounted for as hillslope flow. A true physical model would attempt to account for all of the processes we know occur in firn/crust transport to channels, but since that knowledge (and observations) is not sufficient to parameterize such a model, we use the hillslope friction coefficients as a bulk control on the speed of water as it moves over/through the ice/crust/firn and into channels. We now state the following to avoid confusion "That is, since we lump all flow over/through the ice/firn/snow/crust before it reaches channels into a single 'hillslope' flow with a single friction, we can be confident in the speed of this transport but not its flowpaths or mechanism."

Line 420: Can you give an example(s) of the physical processes that may be leading to this to support this conclusion?
We see how this was confusing, as the use of positive/negative sentence construction was poor. We now state more plainly "However, we believe that model calibration statistics at the outlet indicate the physical realism of the process we're attempting to model: since we modelled an accurate outlet hydrograph, the fully mass and momentum conserved physics of HRR mean that upstream flows must be realistically represented or we could not have produced a quality outlet hydrograph."

Line 443: Anywhere or just in the bare-ice ablation zone?

We believe our study can be repeated anywhere there is a stream network, regardless of the glaciologic regime. In practice, stream networks form most often in this zone, but the physics of open channel flow will be the same in all networks across all settings. The parameters will change to yield different flow velocities through hillslopes and channels, but the basic premise will hold.