Reply to Bettina Schaefli (Referee #1)

We would like to thank Bettina Schaefli for the time she has invested in reviewing our manuscript and the many thorough and constructive comments and suggestions, from which the paper has greatly benefited. Please find below our replies to the individual comments (reviewer comments in italics).

- *The idea of visualizing the data set observation scales is very nice. Food for thought: the MODIS / Landsat data seems very complete in this representation even if they represent only a partial aspect of a state variable (absence presence). There is a dimension missing here of the observation scale (match between simulated and observed flux / state variable content). How could this be integrated?*

  This is a very good suggestion – MODIS and Landsat in fact appear like near-perfect validation data sets in the original visualization of the observation scales, despite their usefulness being somewhat limited due to the binary nature of the snow cover products. In order to integrate this match between simulated and observed fluxes/state variable contents into the visualizations, we have added a new dimension to the radar charts, corresponding to the "information type" of the data sets (either binary or continuous). The hexagons from the original manuscript have hence now become heptagons, with the new axis oriented vertically upwards, thus conserving the symmetry of the charts (spatial scales on the left, temporal scales on the right). The manuscript has been updated accordingly.

- *The runoff module has 12 parameters, which is a very high number; it is in particular difficult to justify to have so many different time scales for the fast component of the flow (not going into the soil). The discharge will never contain enough information to identify them. It is also rather unusual to have a constant fraction of water going into the soil and a single time scale for the soil. This should certainly by re-thought for future applications of this model.*

  We agree that the runoff module has a very high number of parameters which probably cannot all be identified from the observed discharge, as well as that its rather simplistic nature is likely the most limiting factor in our model. It was originally developed as a pragmatic approach to be applied in high-elevation, mostly glacierized catchments, where these types of runoff concentration models tend to perform comparatively well. However, as can be seen in our results its performance especially in larger, less glacierized catchments, is not satisfactory. For future applications it is planned to implement an enhanced runoff scheme in the model.

- *Section 3.5: How is the calibration done (algorithm? Manual? How many model runs?) Is this limiting or are you confident to have found a good solution? Could you have presented model ranges for the mass balance / discharge time series rather than the result of a single parameter set?*

The only model parameters that are calibrated are the parameters of the linear reservoir model (as listed in table 2 of the original manuscript). As this does not require re-running the entire model but can be performed in a post-processing step, this is very performance-efficient (calculating hourly runoff for a single parameter set and catchment takes in the order of 100 ms for a 16-year period, whereas running the entire model for all catchments takes approx. 3–4 days on a 4-core i7 CPU). The calibration is performed using the PyGMO library (`http://esa.github.io/pygmo`), using a combination of differential evolution and L-BFGS-B as optimization algorithms, with the aim to maximize the target function

$$\mathrm{NSE}_V = \mathrm{NSE} - 0.1|V_E|. \tag{1}$$

Hence, presenting model ranges for the discharge time series would have been possible, however not for the mass balance results. Fig. 1 exemplarily shows the parameter values for the 50 best-performing (in terms of eq. (1)) parameter combinations for the Rofenache and Gepatschalm catchments, as well as the skill scores NSE, PBIAS, and BE for the validation period 2007–2013. While the NSE and PBIAS values of all combinations are comparatively similar, a remarkably large spread of the of the resulting values for the benchmark efficiency BE can be observed. However, the best-performing parameter sets in terms of eq. (1) (dark green lines) are also among the best-performing sets in terms of BE. Fig. 2 shows the resulting runoff of the best-performing parameter set as well as the range of the 50 best-performing parameter sets for the two catchments and the year 2012. We have added this information also to the revised manuscript.

- *Areal precipitation: the variation of the soil water storage cannot simply be neglected at the monthly scale; what about ground water? A short comment would be useful.*

  Mean annual catchment total precipitation is calculated from measured runoff, water storage in terms of glacier mass changes known from glacier inventories, and evaporation. Interannual changes in ground water storage are neglected, as the method aims at mean annual values for multi-year periods. The catchment total precipitation is distributed over the months using the measured monthly share on annual precipitation, after which the monthly catchment mean runoff originating from snow melt, ice melt and precipitation minus evaporation is compared to measured runoff. These monthly residuals are related to monthly differing liquid water storage and release within the snowpack and the ground. The monthly liquid storage values are restricted, since they have to feed the base flow in winter and to equalize over the year. For more details please refer to Kuhn (2000, 2003) and Kuhn et al. (2016). We have added this information also to the updated version of the manuscript.

- *How does the uncertainty on initial ice thickness distribution impact the results?*

  We performed some simple sensitivity tests regarding the initial ice thickness distribution by altering the original ice volume by ± 30 % for all glaciers. The results show that for the time scales investigated in our study the initial ice volume has a rather small effect on the results: the mean regional-scale mass balance 1997–2006 changes by -7 to $+6$ %, whereas the cumulative specific mass balance for HEF, KWF, and VF for the period 1997–2013 is affected by -3 to $+6$ %. When performing multi-decadal model runs (e. g., scenario simulations) the influence of the initial ice thickness distribution on the mass balance and runoff results is expected to be considerably larger, which is an effect we will investigate in more detail in our future studies.

- *Section 5.3.1: This section discusses in detail the differences between MODIS and Landsat (which is very useful) but lacks a concise discussion of / conclusion on what these data sets say (respectively can say) about how good the model simulates snow accumulation. As far as I see the model does a good job in winter, when the data does not contain a lot of information (there is snow everywhere) but a bad job in autumn / spring.*

  This is correct – the lowest model performance is obtained during the summer months (as indicated by the Landsat comparisons), but model performance during fall and (to a slightly lesser extent) spring is also not entirely satisfactory. This is mostly due to frequent overestimations of the total snow-covered area, especially at low elevations – possibly due to too high snow correction and snow redistribution factors applied in these areas. Fig. 3 shows the model performance with regard to MODIS separately for fall (OND), winter (JFM), spring (AMJ), as well as these three combined, as a function of elevation. We have added this figure to the supplementary material and added one concluding sentence to the MODIS/Landsat results discussion in the manuscript:

  "To summarize, these results show that the total snow-covered area is frequently overestimated, on average by 20 % (60 %) for MODIS (Landsat). Whereas in only 1 % (0 %) of all investigated scenes snow cover is underestimated by more than 5 %, it is overestimated by more than 5 % in 51 % (92 %) of all scenes. The largest mismatches between the observations and the simulations occur during the summer months (where accurately reproducing the snow cover is the most challenging), followed by fall and spring. Elevation-wise the largest errors occur below the lowest-elevated runoff gauge (further analyses are shown in the supplementary material). A likely explanation for this effect (which is also observed in the snow depth comparisons) is that snow correction and snow redistribution factors are overestimated in these low elevations."

- *It could be interesting to discuss the extreme year 2003 for the mass balance simulations; why is the model particularly wrong here?*

  The simulated mass balances for the three glaciers with annual measurements are indeed particularly underestimated in the year 2003. Looking at the satellite comparison using Landsat and MODIS data for this year (fig. 4), model performance is indeed decreasing during the summer of 2003, however as the BIAS results and the plots of the two Landsat scenes in late July and mid-September, respectively, show (figs. 5–6), this is due to an over- rather than an underestimation of snow cover (at least on these two dates). Runoff simulations (fig. 7) for the gauge Vent/Rofenache (of which the three glaciers are contributing to) also show no striking anomalies, but rather a quite satisfying agreement with the observations during the summer months of 2003. After further investigation, comparison of simulated snow depth for three stations (fig. 8), showed that for the stations Obergurgl and Pitztaler Gletscher snow depth is captured very well (Obergurgl) and slightly overestimated (Pitztaler Gletscher), respectively, however distinctly underestimated for the station Vernagtbach (being the closest station to the three investigated glaciers, hence having the highest influence on the interpolated meteorological fields) in this year, mostly due to a precipitation event in mid-November 2002 which is not captured by the precipitation recordings at this station (blue line in fig. 8). Re-running the model while excluding the precipitation recordings for October and November 2002 at this station (i.e., precipitation for the station location is then obtained by interpolation from the surrounding stations) yields considerably higher snow depths (green line in fig. 8). Fig. 9 shows the cumulative daily glacier mass balances for HEF, KWF, and VF for the glaciological year 2003. The simulated mass balances are still underestimated, but approx. 250 mm (HEF), 450 mm (VF), and 500 mm (KWF) less negative than in the original model run, indicating that the deviations in simulated mass balance for this year can at least partly explained by this issue. We have included these results in the supplementary material and briefly discuss this issue in the updated manuscript.

- *P. 17, line 6: it is stated here that there is no calibration on the glacier mass balance; this statement should be re-thought in light of the fact that there is a snowfall correction factor calibrated on all available glacio-hydrologic data; I would also not include areal-precip in the set of validation data since it is actually used to estimate a key parameter, not for validation (reformulate abstract and conclusion)*

  It is correct that the snow correction factor has been determined using the glacio-hydrological data, however these analyses were performed on the regional scale and not on the mass balances of the three individual glaciers. We have reworded the sentence in question accordingly:

  "The generally satisfying model performance for the three glaciers indicates that the model setup is suitable for glacier mass balance simulations at the regional scale, as no glacier-specific model calibration has been performed."

  With regard to the second comment, it is true that areal precipitation is used to estimate the additional snow correction factor, however only at the scale of the entire study site (all catchments). Validation against areal precipitation is on the other hand performed for the individual catchments. Hence, the model is calibrated against the total precipitation volume, but validated against its spatial distribution. While areal precipitation thus cannot be considered a truly independent validation source, in our opinion it is still valid to include it in the list of validation data sets.

- *Does the implementation of the cold content / liquid water module for the snowpack improve discharge simulations? If yes: I would show it (perhaps in the supplementary material), otherwise: why did you choose to implement it?*

  The long-term discharge simulations were not improved due to the implementation of the module for cold content and liquid water content, as can be seen in table 1 (skill scores for the individual catchments and the validation period 2007–2013 obtained using individually calibrated runoff parameters for the period 1997–2006) – in fact average model skill even decreased slightly after implementation of the module. The module was rather implemented to improve the snow cover simulations and prevent premature melting of snow in high elevations. As table 2 shows, snow depth evolution was only slightly improved at the high-elevated stations, while at the low-elevated stations snow depth overestimations increased even more, resulting in lower model performance. However, the implementation of the cold content/liquid water content module drastically improved the simulations regarding (i) snow depth distribution over the winter 2010/11 (fig. 10), (ii) long-term glacier mass balance 1997–2006 (fig. 11), and (iii) annual (fig. 12) and cumulative (fig. 13) glacier mass balance for HEF, KWF, and VF, as can be seen when comparing these figures (obtained without consideration of cold content and liquid water content) with the ones shown in the original manuscript. We have added these figures to the supplementary material.

- *Section 5.5: your results suggest that the precipitation amounts are not compatible for glacier mass balance and discharge (discharge bias increase for corrected precipitation). This is not unusal but could you comment on it? Could the bias be explained by errors in other water balance terms (evapotranspiration, groundwater exchange)?*

The effect of increased discharge bias for corrected precipitation might partly be attributed to overestimated precipitation (e. g., for the Rofenache catchment – where runoff volume after correction is overestimated by approx. 20 % in the validation period – AMUNDSEN mean areal precipitation is 8 % larger than the value derived by closing the hydrologic balance), however this cannot be observed in all catchments (e. g., for the Gurgler Ache catchment, precipitation after correction is still lower than the hydrologic balance-derived precipitation, yet runoff bias strongly increases). In the catchment with the highest glacierization (Pitze), where the applied precipitation correction leads to the highest precipitation increase of all catchments (650 mm), the effect is much less pronounced (runoff bias in the validation period only increases from 2.9 to 4.4 %). A comparison of the simulated evapotranspiration values with the respective values derived using the OEZ method as well as with other studies in the Ötztal area (e.g., Tecklenburg et al., 2012; Kormann et al., 2016) indicate that evapotranspiration might be underestimated by the model. Hence, it is in fact likely that the runoff volume bias can at least partly be attributed to other water balance terms (evapotranspiration and groundwater exchange). At the moment the water balance model WaSiM is being set up for the study site as part of the project HydroGeM3 (funded by the Austrian Academy of Sciences). As the methods for the estimation of evapotranspiration and groundwater fluxes implemented in WaSiM are clearly superior to those in AMUNDSEN, through intercomparison of the model results we aim to get a deeper understanding of the causes for the discharge overestimation. However, unfortunately no results are available yet to be added to the manuscript.

- Minor comments (*no need to answer in the public discussion*): All of these comments were considered in the preparation of the revised manuscript, except for the last one (*Figure 14: I would represent also storage changes and evapotranspiration*). As with this figure we intended to show solely the fractions of the individual runoff components to total runoff, we have left the figure unchanged for now. Presenting the total water balance of the catchments including storage changes and evapotranspiration would of course be interesting, however would require further analyses as well as preparing an additional figure. However, if the editor deems this appropriate, we would be happy to do so.

References

Kormann, C., Bronstert, A., Francke, T., Recknagel, T., and Graeff, T. (2016). Model-Based Attribution of High-Resolution Streamflow Trends in Two Alpine Basins of Western Austria. *Hydrology*, 3(1):7.

Kuhn, M. (2000). Verification of a hydrometeorological model of glacierized basins. *Annals of Glaciology*, 31(1):15–18.

Kuhn, M. (2003). Redistribution of snow and glacier mass balance from a hydrometeorological model. *Journal of Hydrology*, 282(1-4):95–103.

Kuhn, M., Helfricht, K., Ortner, M., Landmann, J., and Gurgiser, W. (2016). Liquid water storage in snow and ice in 86 Eastern Alpine basins and its changes from 1970–1997 to 1998–2006. *Journal of Glaciology*.

Tecklenburg, C., Francke, T., Kormann, C., and Bronstert, A. (2012). Modeling of water balance

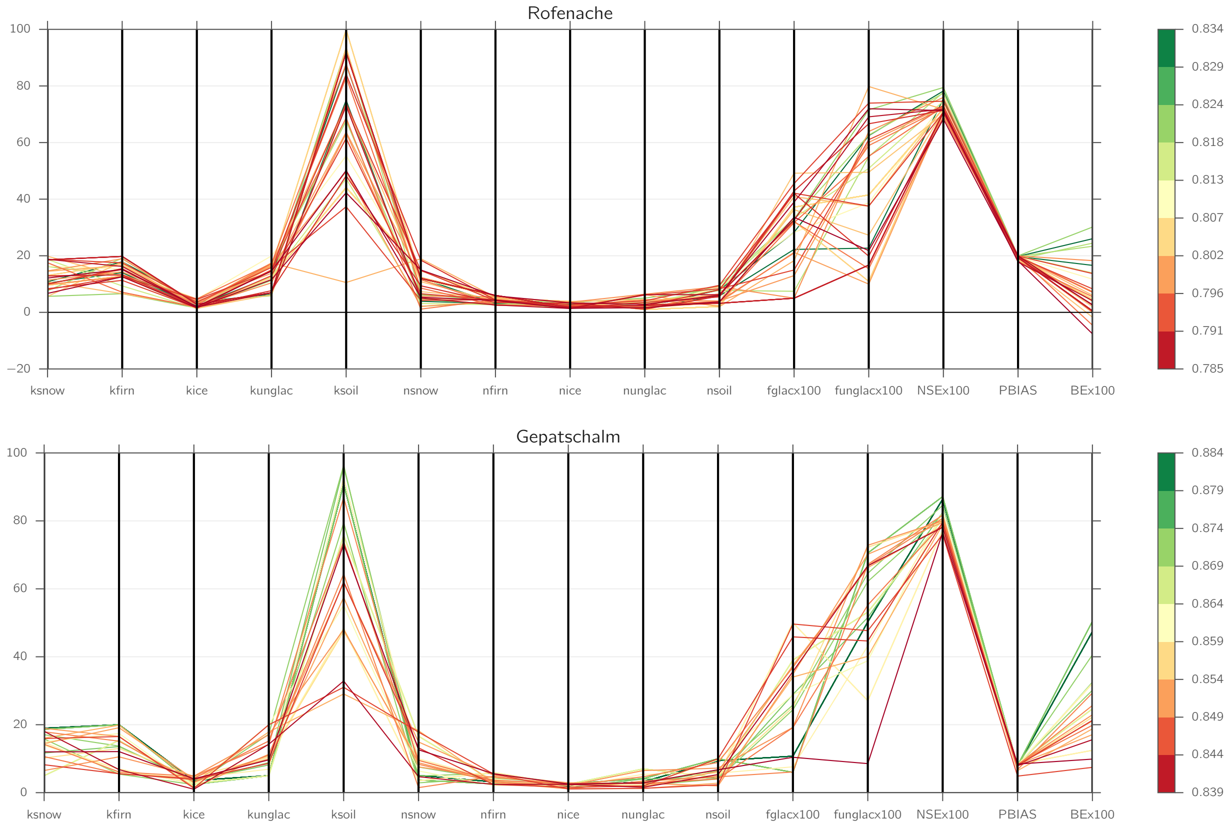

Figure 1: Parameter values for the 50 best-performing (in terms of eq. (1)) parameter sets, as well as the skill scores NSE, PBIAS, and BE obtained for the validation period 2007–2013, for the Rofenache (top) and Gepatschalm (bottom) catchments. Line colors correspond to the calibration period (1997–2006) model skill in terms of eq. (1).

response to an extreme future scenario in the Ötztal catchment, Austria. *Advances in Geosciences*, 32:63–68.

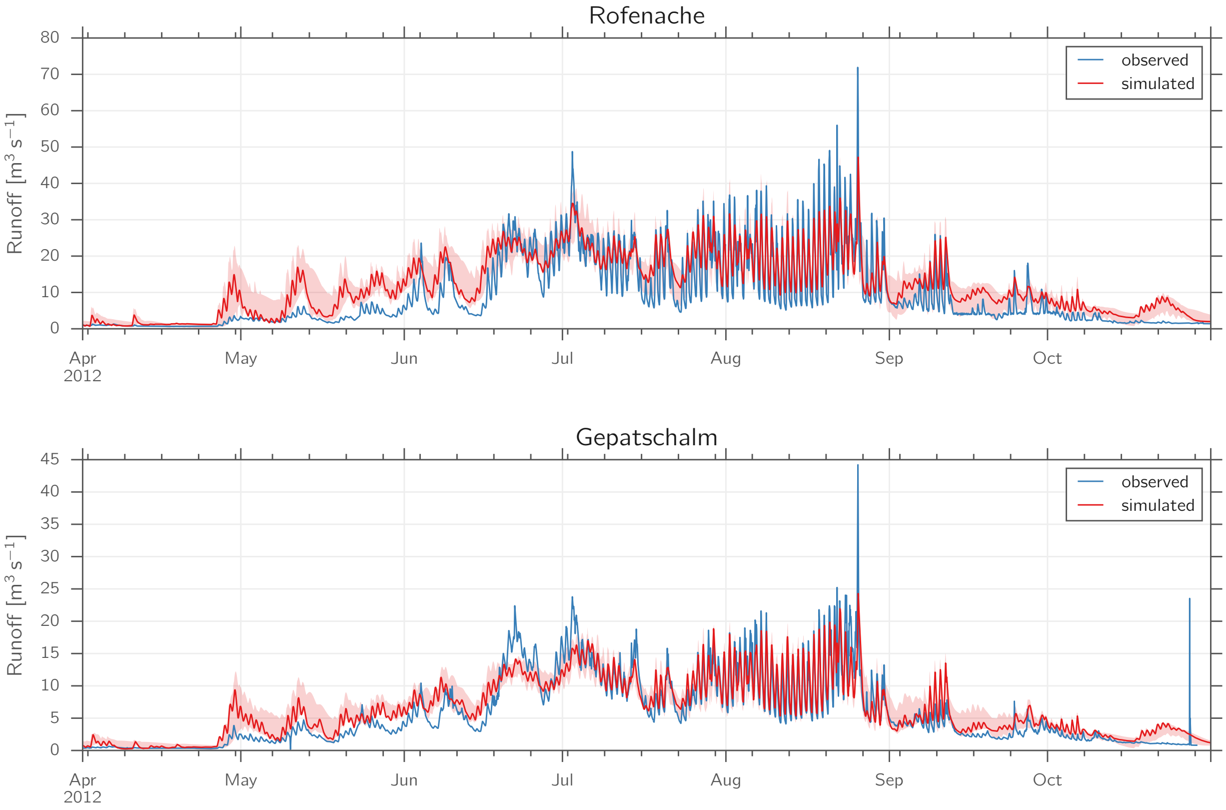

Figure 2: Observed (blue) and simulated (red) runoff for the Rofenache (top) and Gepatschalm (bottom) catchments in the year 2012. The red lines correspond to the runoff resulting from the best-performing parameter set in the calibration period, whereas the shaded areas represent the range of the 50 best-performing parameter sets.

Table 1: NSE, BE, and PBIAS of simulated vs. observed runoff for the validation period 2007–2013 obtained using model runs without and with consideration of cold content and liquid water content, respectively.

		Without CC/LWC			With CC/LWC		
ID	Catchment	NSE	BE	PBIAS	NSE	BE	PBIAS
1	Rofenache Am	0.78	0.35	20.0	0.78	0.35	20.0
2	Barst/Gurgler Ache	0.80	0.26	19.1	0.80	0.26	20.8
4	Gepatschalm	0.87	0.48	6.9	0.87	0.50	7.3
6	Taschachbach	0.88	0.45	0.2	0.88	0.45	2.2
7	Pitze	0.87	0.53	4.7	0.87	0.53	4.4
8	Radurschlbach	0.67	-0.39	20.3	0.67	-0.43	22.7
13	Fissladbach	0.67	-0.14	28.8	0.63	-0.23	31.9
	Mean	0.79	0.22	14.3	0.79	0.20	15.6

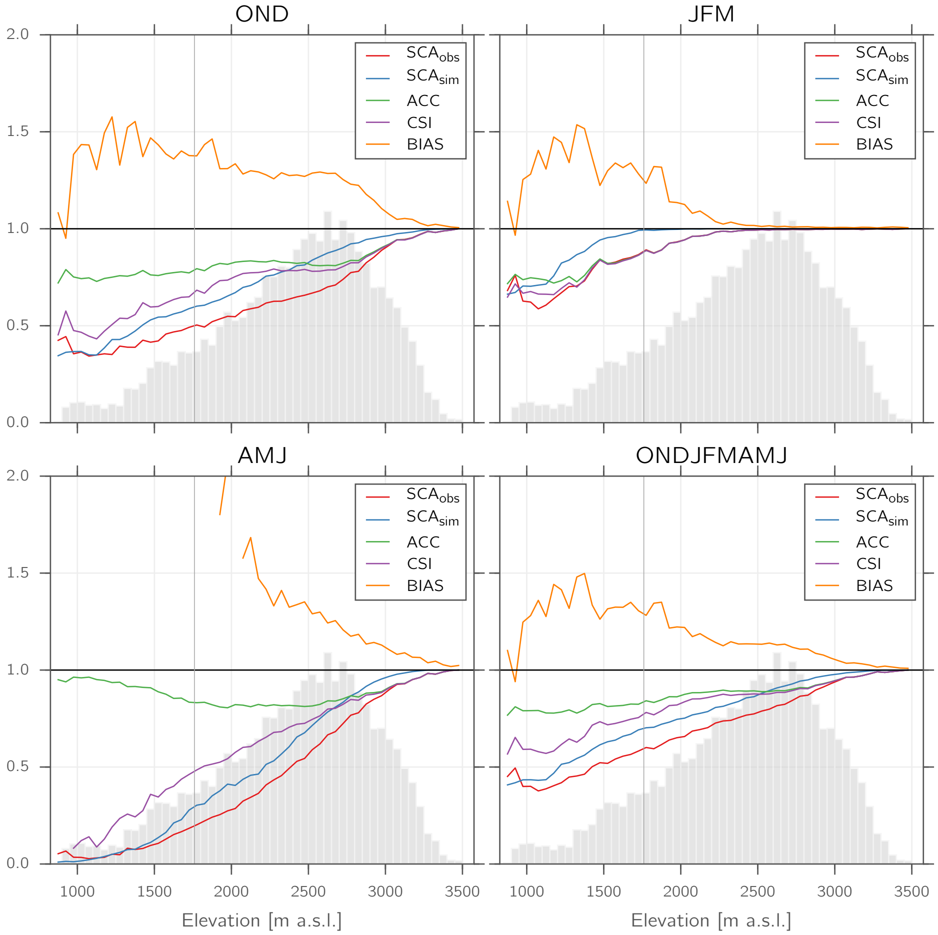

Figure 3: Average values of SCA, ACC, CSI, and BIAS by elevation of all suitable MODIS scenes shown for Oct–Dec (top left), Jan–Mar (top right), Apr–Jun (bottom left), and Oct–Jun (bottom right). The vertical lines indicate the elevation of the lowest-elevated runoff gauge at 1760 m a.s.l., and the gray bars represent the relative area distribution of the elevation bands.

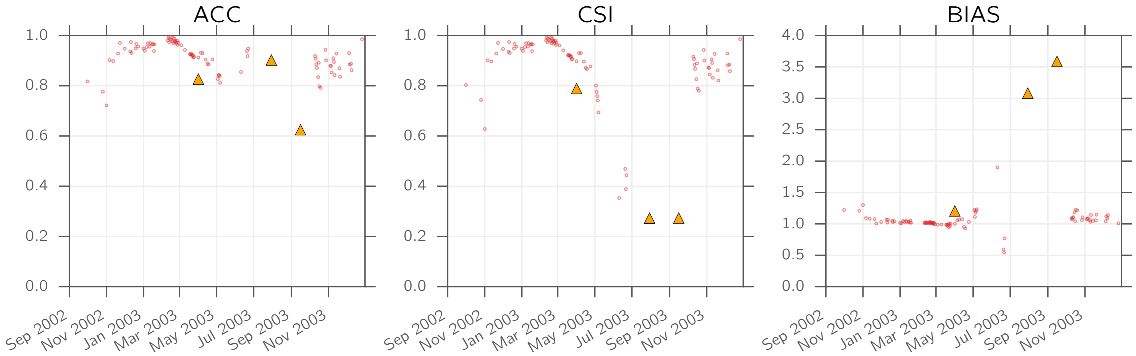

Figure 4: ACC, CSI, and BIAS for all available Landsat (triangles) and MODIS (circles) scenes of the year 2003.

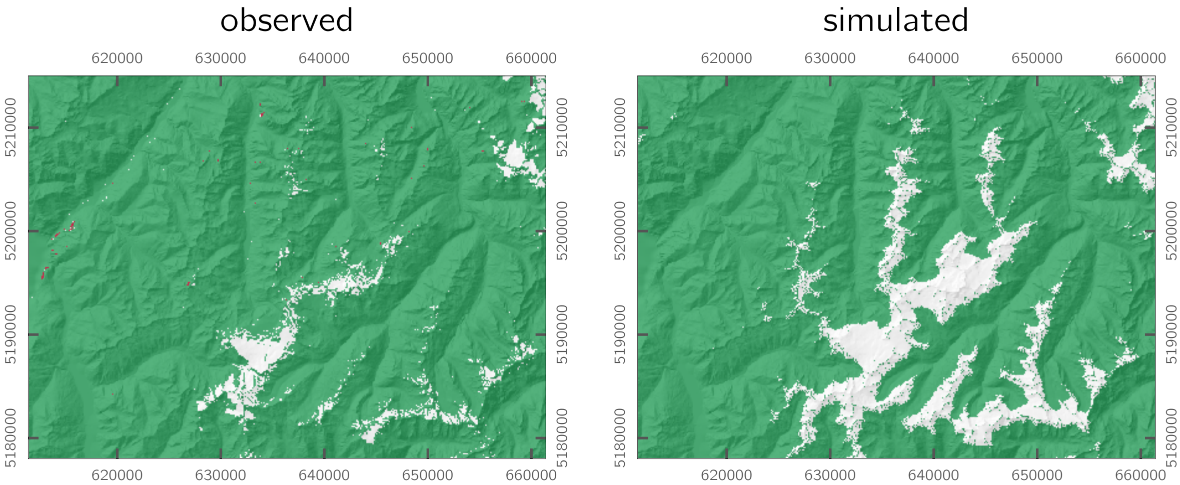

Figure 5: Observed (Landsat) and simulated snow cover distribution for July 30, 2003.

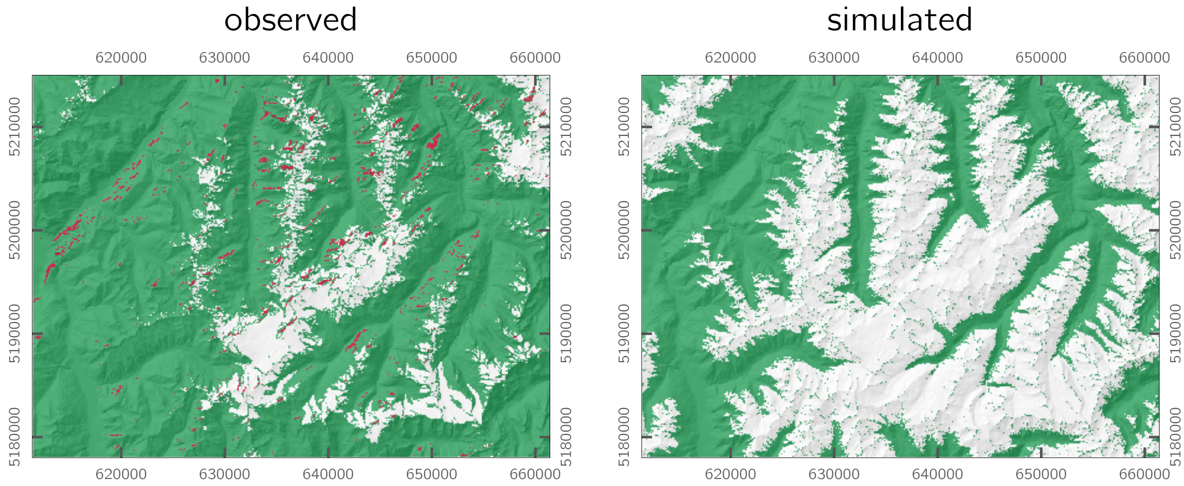

Figure 6: Observed (Landsat) and simulated snow cover distribution for September 16, 2003.

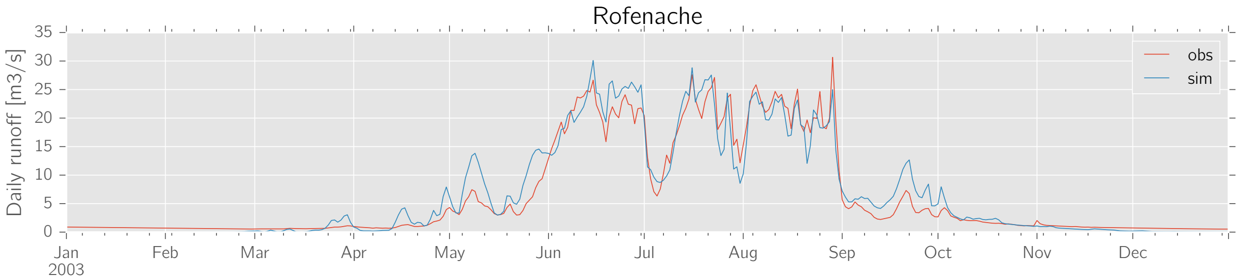

Figure 7: Observed and simulated mean daily runoff for gauge Rofenache (including HEF, KWF, and VF glaciers) in the year 2003.

Table 2: R^2, Nash-Sutcliffe efficiency NSE, and percent bias PBIAS for observed vs. simulated snow depth obtained using model runs without and with consideration of cold content and liquid water content, respectively.

		Without CC/LWC			With CC/LWC		
Station	Elevation [m a.s.l.]	R^2	NSE	PBIAS [%]	R^2	NSE	PBIAS [%]
Prutz	871	0.70	-1.09	170.87	0.66	-2.21	222.5
Nauders	1330	0.79	0.04	99.02	0.74	-0.35	119.5
Obergurgl	1942	0.76	0.74	-13.32	0.85	0.85	-6.0
Weisssee	2480	0.69	0.57	13.45	0.76	0.52	28.7
Pitztaler Gletscher	2864	0.74	0.72	-9.54	0.71	0.68	7.1

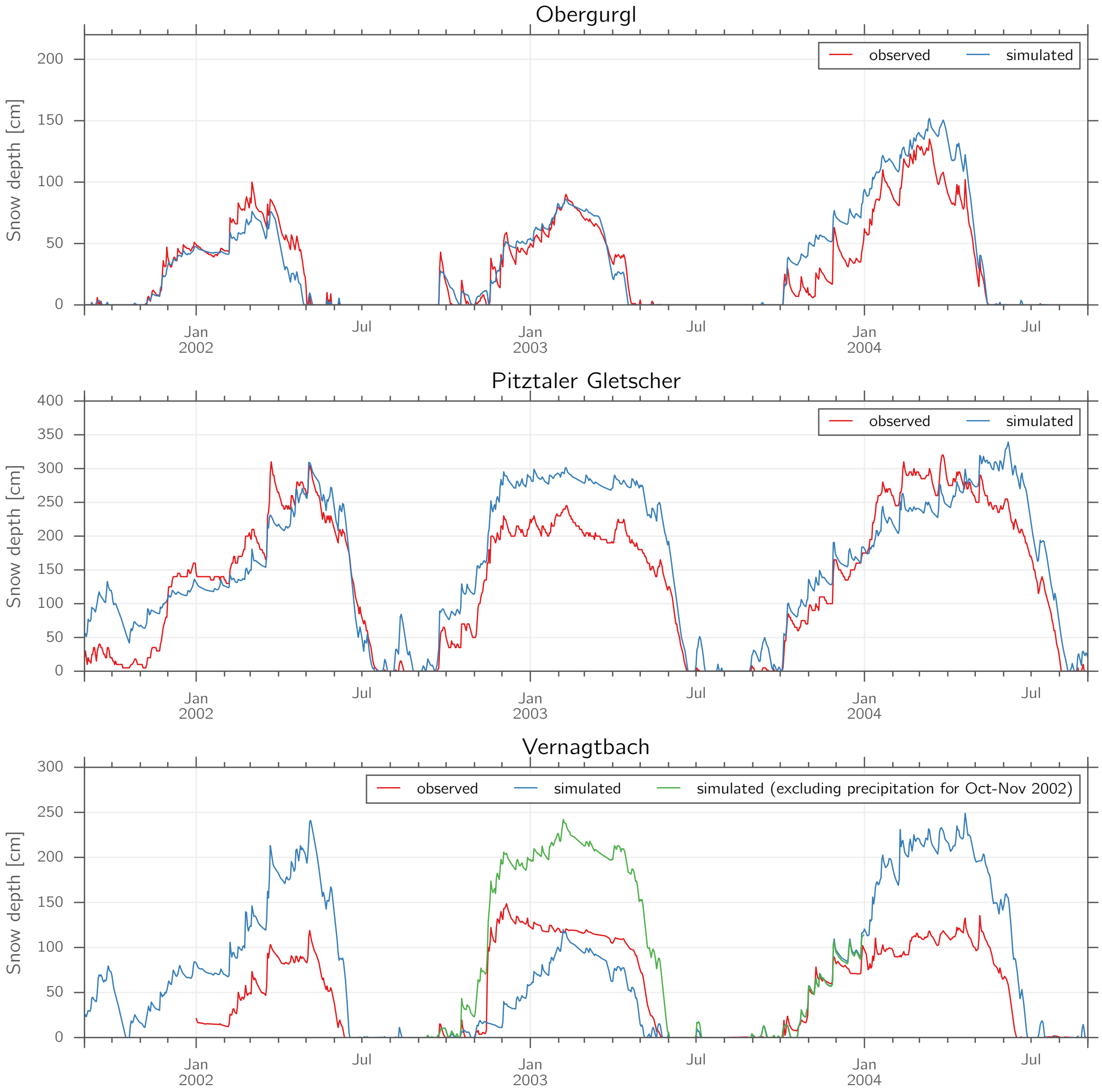

Figure 8: Observed and simulated snow depth for the stations Obergurgl (1942 m a.s.l.), Pitztaler Gletscher (2864 m a.s.l.) and Vernagtbach (2640 m a.s.l.) in the period 2001–2004.

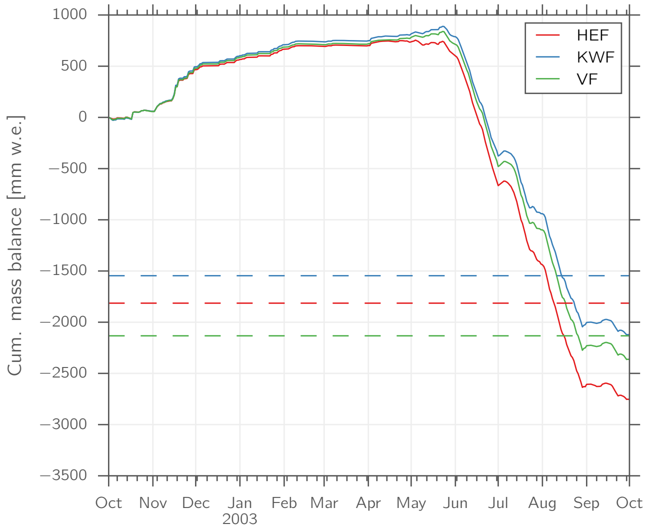

Figure 9: Cumulative daily mass balance for the glaciological year 2003 and the glaciers HEF, KWF, and VF as simulated by AMUNDSEN while exluding the Vernagtbach station precipitation recordings for October–November 2002 (solid lines), and observed glaciological mass balances for this year (dashed lines).

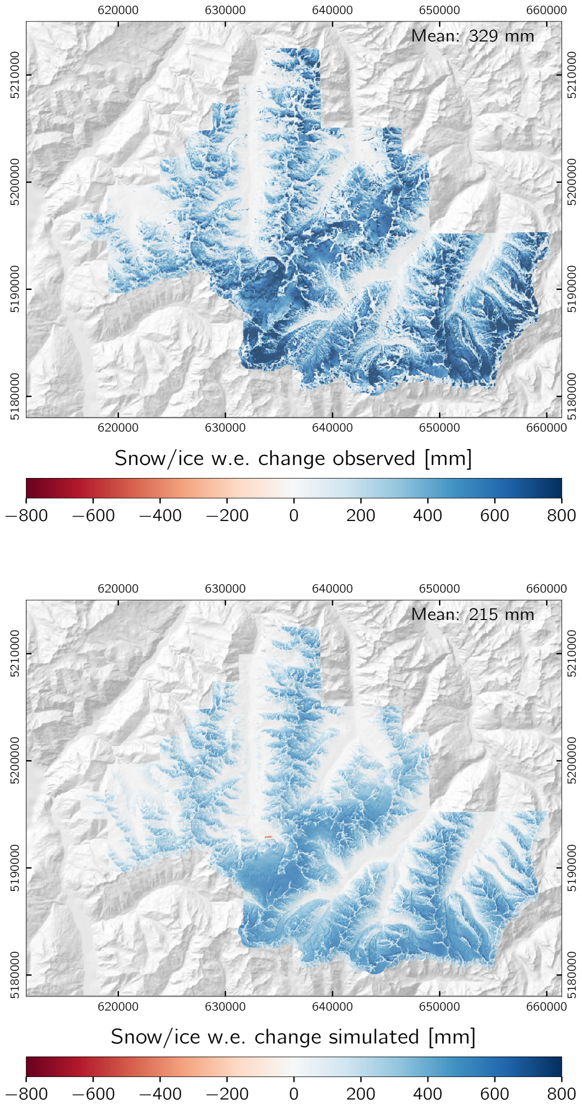

Figure 10: Observed and simulated end-of-season snow distribution for the winter 2010/11 (October 8 to April 22) without consideration of cold content and liquid water content in the model.

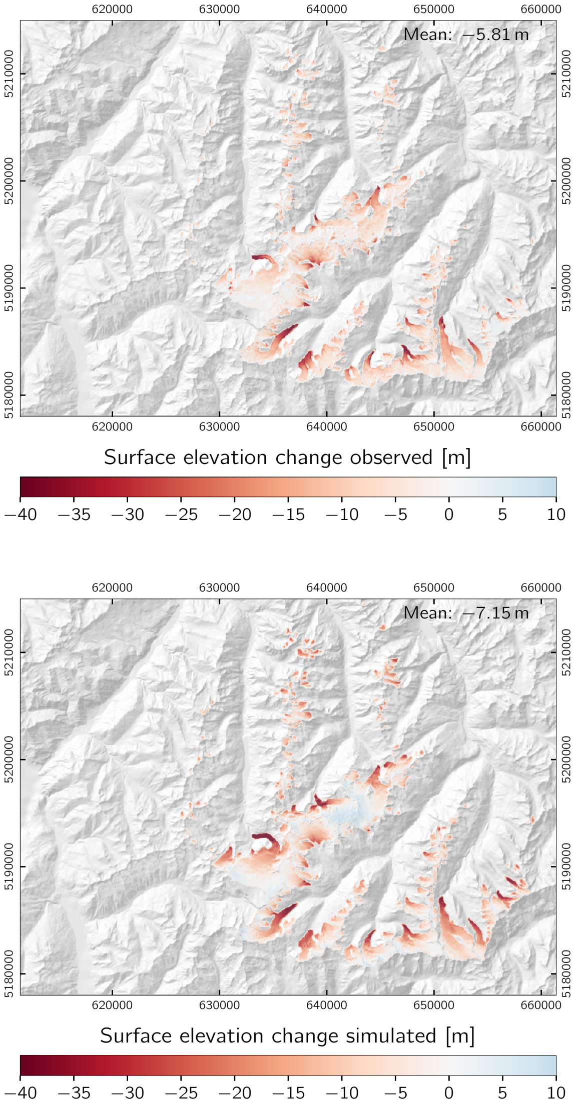

Figure 11: Observed and simulated glacier surface elevation change for the period 1997–2006 without consideration of cold content and liquid water content in the model.

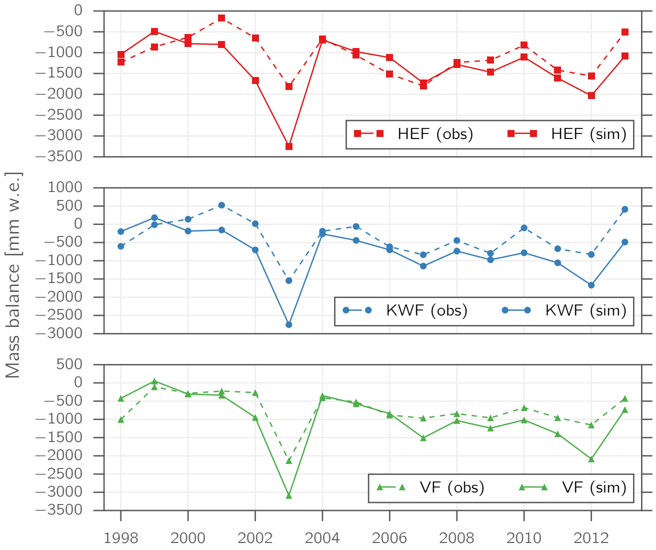

Figure 12: Observed (dashed) and simulated (solid) specific mass balance for Hintereisferner (HEF, top), Kesselwandferner (KWF, center), and Vernagtferner (VF, bottom) in the period 1997/98–2012/13 without consideration of cold content and liquid water content in the model.

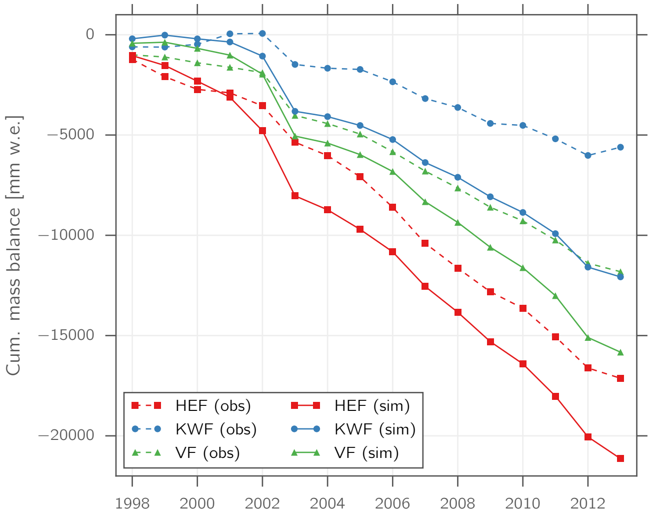

Figure 13: Observed (dashed) and simulated (solid) cumulative specific mass balance for Hintereisferner (HEF), Kesselwandferner (KWF), and Vernagtferner (VF) in the period 1997/98–2012/13 without consideration of cold content and liquid water content in the model.