Reply to the reviews on our manuscript "ENSO influence on surface energy and mass balance at Shallap Glacier, Cordillera Blanca, Peru"

F. Maussion, W. Gurgiser, M. Großhauser, G. Kaser, and B. Marzeion
Institute of Meteorology and Geophysics, University of Innsbruck, Innsbruck, Austria

We would like to thank Tobias Bolch for obtaining the reviews and both anonymous reviewers for providing fast, detailed, and helpful comments on our manuscript. We reproduce the reviewers' comments below and provide our answers in italic.

The changes made in the manuscript are summarized here:

- *ENSO classification is now monthly and not based on hydrological years any more. This has only a small impact on our results but it clarified the signal found in the composites.*
- *we strengthened the discussion about the glacier-wide mass-balance and its uncertainty. We removed the reference to the average absolute mass-balance values in order to emphasize that the focus of our manuscript and of the chosen methodology is put on variability, not the exact MB values.*
- *we added a new appendix figure where we discuss the SSTA → MB relationship for the longer 1950–2013 period, based on NCEP/NCAR data.*
- *we added a new paragraph to the sensitivity analyses, where we compare various ENSO indicators (Nino 1+2, Nino 3.4 SSTs, and MEI) and various lags. The differences between the indicators are small and well below the uncertainty ranges of the modelled MB.*

Response to Anonymous Referee #1

{ Introduction text removed }

Comments/Questions:

- I dont understand why the ENSO definition has to be based on full hydrologic years. After all the analysis was based on monthly time steps, which would have allowed for a much more refined ENSO delineation. Allowing a full year to be counted as El Niño or La Niña year as long as 5 out of 12 consecutive months are above or below the threshold does not seems like a very stringent criterion. Indeed when looking at Figure 2 it look as if some years, which were essentially neutral years were classified as La Nina. I think the composites of the seasonal cycle associated with the two phases of ENSO shown in Figure 6 would be much cleaner as a result of a better ENSO phase definition.

This choice was motivated by the fact that we are also interested in ENSO-related shifts of the annual cycle. This allowed us to build "consistent" composites of the annual cycle (Figs. 6 and 7) with the same number of values for each month.

However, we agree with the arguments of both Reviewers and changed our analysis accordingly. The Niño / Niña periods are now defined exactly according to Trenberth (1997) (Fig. R1 below). The same periods are shifted forward by three months to account for the lag between MB and SST and then used to build the new composites. The difference with previous results is quite *small, but it is best shown with the updated Fig. 6 (Fig. R2 below). We also changed Fig. 7 in the manuscript accordingly.*

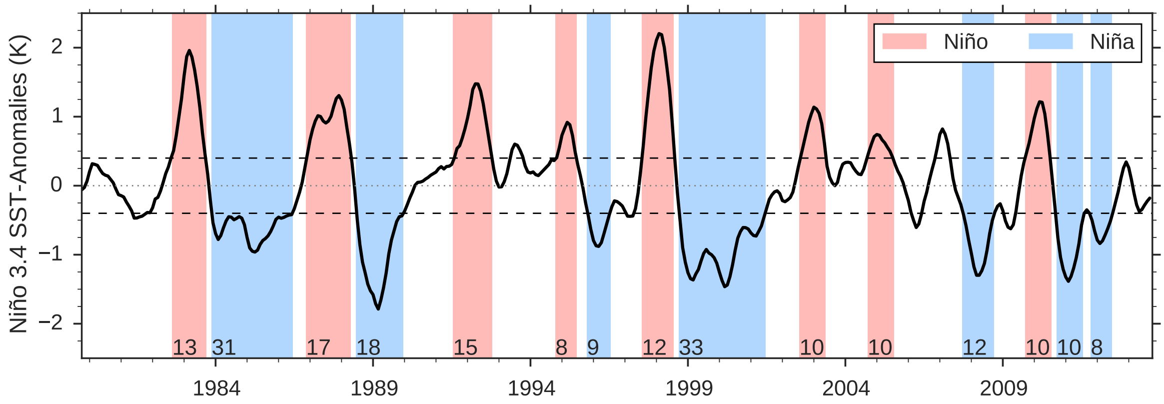

Figure R1: Updated Fig. 2 with the new ENSO classification. The length of each period is indicated at the bottom of the plot.

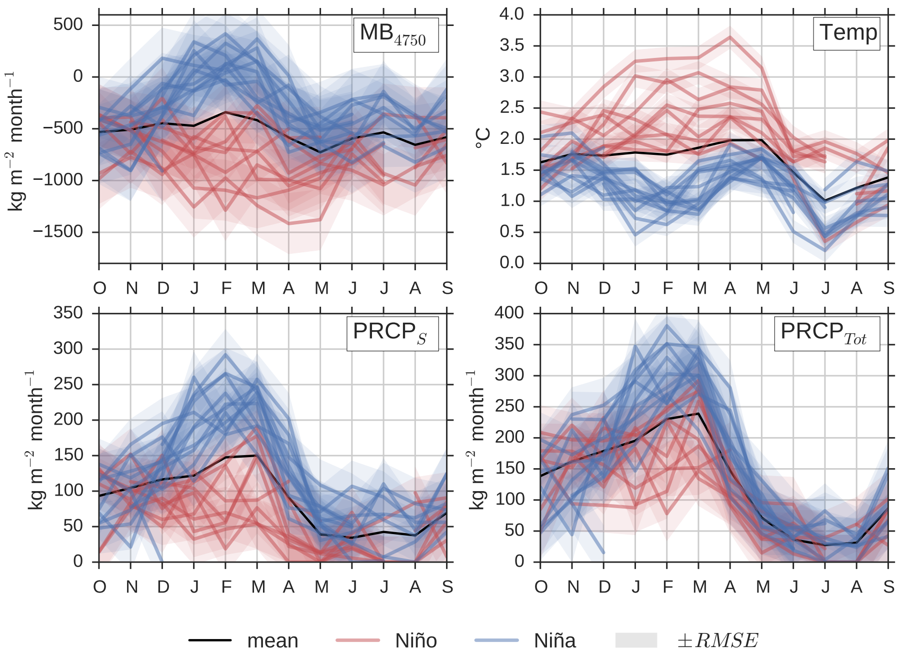

Figure R2: Updated Fig. 6 with the new ENSO classification. Note that some annual cycles are now incomplete.

- The atmospheric circulation in the region undergoes a fundamental seasonal transition from wet to dry periods. I wonder if this does not pose significant restriction on downscaling methodology (i.e. would results be different if separate downscaling algorithms were used for wet, dry and transition seasons?).

This is a valid concern and indeed, there are arguments in favour of "seasonally dependent" models. First, a model calibrated/validated over entire years will have a facilitated job because *of the presence of an annual cycle in some of the predictands/predictors, leading to an overestimation of the model's skill. This led authors such as Hofer et al. (2010) to define a separate model for each month. Second, it is probable that the large-scale drivers of precipitation are changing for various seasons, which calls for more flexible models.*

This second argument, however, can also be turned against seasonal models. It could happen, for example, that "March-like" circulation occurs exceptionally in May that year. A model calibrated from January to March during a reduced number of years might not have been prepared for this situation, while an annual model might be. This is particularly true in the Cordillera Blanca, where the length of the wet season varies greatly and where several variables have a very low seasonality. It is a challenge to define a period over which the seasonal model should be calibrated.

That said, it is not possible for us to test which model would work best: the period of available data at hand is too short. The fact that the downscaling model is able to produce wet seasons of varying lengths as result of an atmospheric forcing is an indication for a certain flexibility. The skill score used in the paper to test the model should also partly account for the issue of the annual cycle, since the reference model (simple climatology) also has the same "advantage" as the downscaling model.

- The downscaling model only accounts for local relationships with the large-scale circulation (i.e. the closest reanalysis grid cell). The authors argue that this is justified as it ensures to allow for the local climatic influence and avoids spurious long distance in- fluences that may not be real. I agree that this is a valid argument, but at the same time there are dynamical reasons why the strongest relationship with atmospheric variables may not be located directly overhead. For example correlations with an oscillatory mode will almost never be strongest directly aloft and in fact may be completely missed if the location in question is near the node of the oscillation, far removed from the two oscillatory poles. In addition Vuille et al. (2008b) documented that correlation of Cordillera Blanca mass balance with atmospheric temperature is significantly stronger toward the equator than directly overhead (see their Figure 6). Hence I think the down- scaling model could still be improved in future studies by also allowing for more distant influence factors.

We agree. Longer calibration periods would allow the search for more predictive distant variables, such as has been done by Shea and Marshall (2007) and Vuille et al. (2008b). At this stage and for such a short period, the risk of spurious correlation is however very high. Recalling the previous discussion about seasonal models: arguably, distant predictors are likely to have a predictive skill for certain seasons only (pacific SST during austral summer only for example). In turn, the local grid point's atmospheric state (moisture content, temperature...) is more likely to be related to the glacier surface conditions at all times of the year.

Minor edits:
- Table 1: please replace the terms longitudinal and latitudinal wind component with the appropriate meteorological terms: zonal and meridional wind component.

Done.

-Figure caption 8. You state that Pacific SST are lagged by three months. I assume this is a typo, since Pacific SST should obviously lead the mass balance series by three months.

Yes, this can be confusing. Pacific SST leads MB by three months, so one has to "delay" the time series by three month for maximal correlation with MB. We replaced the word "lagged" *with "shifted forward".*

Page 3000, line 13 and throughout manuscript: capitalize pacific
Page 3003, line 14: change Francou, 2003 to Francou et al., 2003
Page 3019, footnote 3: change he to the

Done.

Page 3021, line 1: To my knowledge Salzmann et al (2013) never stated that precipitation is increased in the Cordillera Vilcanota during El Nino periods. Please clarify where they made such a statement or remove this reference.

Correct. Salzmann et al. (2013) found no clear ENSO influence on the region but made no statement about precipitation (only about temperature). We removed the reference but kept the reference to Perry et al. (2014) who indeed found a Niño-wet / Niña-dry signal (their study covers a rather short period of time, however)

Page 3021, line 15: change specially to especially
Page 3025, line 22: change participated to to participated in
Page 3028, line 16: Reference Francou, 2003: you forgot to list the co-authors of that study.
Page 3028, line 17: chacaltaya should be capitalized.
Page 3028, line 19: andes should be capitalized.

Done.

Response to Anonymous Referee #2

{ Introduction text removed }

1. I am wondering how dependent the results are on the choice of the ENSO classification. In this paper, the authors are considering the Niño3.4 region SSTA with a threshold of 0.5K or -0.5K following Trenberth (1997) to separate Niño, Niña or neutral hydrological years. Then part of the analysis (Fig. 6, 7 and text) is done based on this classification. But would the results have been changed using a different threshold (sometimes the used threshold is 0.4K / -0.4K), or the Niño1.2 region (like in Francou et al., (2003) - see specific comment # 9)? It is also surprising to make this classification on an annual basis although the analysis is conducted at monthly scale. Using an annual scale explains why the classification applied in this study differs from the classification obtained with the multivariable ENSO index. I suggest to change the ENSO classification and to follow the classification obtained with MEI. If the authors decide to keep the original classification, some justifications are needed and a comparison of the results using different Niño/Niña periods could be interesting.

The second part of this question is related to the first comment of reviewer # 1. We agreed to your suggestion and now use a monthly classification.
For the dependency to the choice of the SST products, we refer to our response to specific comment # 9, where we introduce new sensitivity analyses with various SST products and lags.

2. I believe that the analysis of the performance of the model especially regarding its ability to reproduce the glacier-wide MB (p3016 mainly) should be done more carefully (see my specific comment # 11). I suggest to add a figure with the

modelled gradient of MB as a function of altitude (VMBG), and to perform a detailed comparison with available observations: comparison between modelled and VMBG measured on some glaciers in the Cordillera Blanca and comparison between the mean modelled MB over the period 1980-2013 and observations (see Rabatel et al. (2013).

We agree that the glacier-wide MB is subject to much larger uncertainties, originating from both the calibration SEB/SMB time series and the downscaling model. There is no information about accumulation at higher altitudes and no way to validate both models.
We plot the vertical mass-balance profiles in Fig. R3 below. One can see that the downscaling model reproduces well the reference mass-balance gradient. This is a fundamental property of statistical models which preserve the average of the target variables. The MB during the reference period appears to be above the average of the 1980–2013 period. The gradients are very steep in the ablation area, and the accumulation area ratio (AAR) of Shallap is approx. 75% (a known feature of tropical glaciers, e.g. Kaser and Osmaston, 2002). In their shape and magnitude, the gradients are similar to those measured at Zongo glacier for example (Sicart et al., 2007, their Fig. 4).

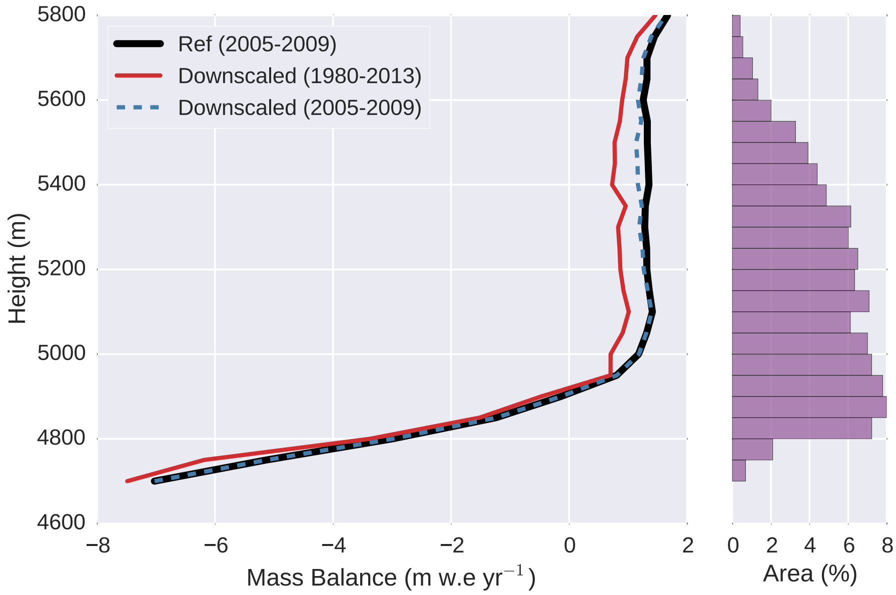

Figure R3: **Left**: averaged mass-balance profiles ($m\ yr^{-1}$). **Right**: area (%) of each altitude slice.

The computed specific MB of -0.04 ± 0.4 *m w.e. for the 1980–2013 should be read with extreme caution. First, the glacier had a larger ablation area during this period, and with such steep gradients the mass loss in the earlier years will be significantly larger on average. In a very simple idealized experiment, we set the area of the lowest elevation band (based on 2001 glacier extend) to +25% in 1980 and let it decrease linearly during 1980-2001 and remain constant for 2001-2013 (where we have no information about its change). This results in a more negative specific MB of* -0.48 *m w.e., much closer to the regional average (Rabatel et al., 2013). The second issue is related to the statistical model's tendency to preserve the mean during the calibration period: systematic over- or underestimations in the calibration time series will have a durable impact over the entire downscaling period.*

These are the two major reasons why we: (i) only focus on SEB/SMB variability rather than absolute values and (ii) arbitrarily doubled the "glacier-wide" uncertainties for the analyses. To avoid future misinterpretations we strengthened the discussion of uncertainties and removed the sentence mentioning the average specific MB of -0.04 ± 0.4 *m w.e. Thanks for pointing that out.*

3. This analysis spans over a 33-year period, but some reanalysis products used in this study are available before 1980 (section 5.3). It would have been interesting to extend back in time this analysis to check if the relationship between SST and MB is still valid before 1980.

Indeed, NCEP/NCAR R1 is available from 1948 onwards, and the SST data from 1950 onwards: the timeseries are plotted below (Fig. R4). In general, NCEP/NCAR MB does not correlate as well as other reanalyses (cf Table 3 in the manuscript), which is expected since NCEP/NCAR is an older dataset with known deficiencies. However, the correlation remains quite stable throughout the 60 years, with lower correlations for the 1951–1980 period. For the downscaled timeseries (without RMSE, without detrending):

- *1951–2013:* $r^2 = 0.45$ ($p < 10^{-5}$)
- *1981–2013:* $r^2 = 0.58$ ($p < 10^{-5}$)
- *1951–1980:* $r^2 = 0.42$ ($p < 10^{-5}$)

We added these analyses in the revised manuscript as a new appendix.

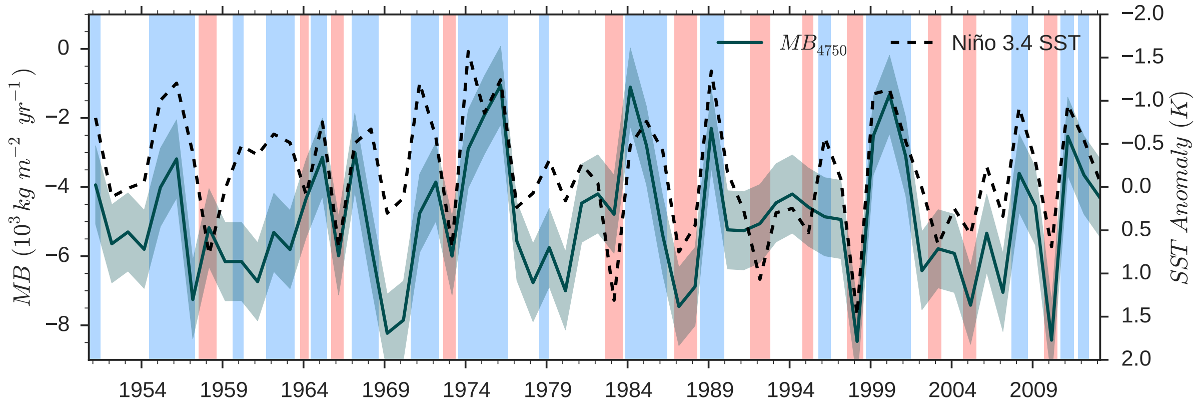

Figure R4: New appendix. Same as the manuscript's Fig. 8 but with NCEP/NCAR reanalysis and for the longer period 1950–2013.

Specific comments

1. P3005, line 1: how can periods overlap when the AWS on the moraine stopped in 2009, and the AWS on glacier started in July 2010?

This precision was missing in the description, now corrected to "Two automatic weather stations were operated over two distinct and partly overlapping periods: at the glacier surface (July 2010–September 2012, with several gaps) and on the southern moraine (2002–2009 and July 2011 to February 2012)".

2. P3005 line 24: it is equivocal to speak about conductive heat flux from the

ground. The energy flux inside snow/ice or into the glacier body is a better formulation

Agreed.

3. P3006 line 15: why assuming that F=QM although in the calculations, surface temperature is probably available? This might bring a substantial bias especially for the elevation slices in the vicinity of the ELA (Fig A1) and also but to a lesser extent in the accumulation area (explaining why the BSS score is low at high elevation as stated p3015, line 21). Moreover, this assumption will limit the transferability of the model to other regions than the tropics i.e. mid-latitude or polar glaciers.

Thanks for this comment. Indeed this is a source of uncertainty, albeit small in comparison to other sources as discussed in Appendix A1. Surface temperature is available from the calibration time series, and is also indirectly downscaled (we can compute it from LW_{out}). However, there is no easy, one-to-one relationship between monthly T_S and the ratio of the energy flux F which is transformed to melt. This is due to many factors such as the time of day when melt occurs, temperature, altitude, etc.

That said, in the early stages of this study we investigated a possible way to reduce this uncertainty, and developed a methodology which is still available as an optional computation of melt in DownGlacier. We introduced a correction factor c which guarantees that the average downscaled melt mass is equal to the average calibration melt mass. The usage of this factor can also be cross-validated, as shown by the scores in Fig. R5 below.

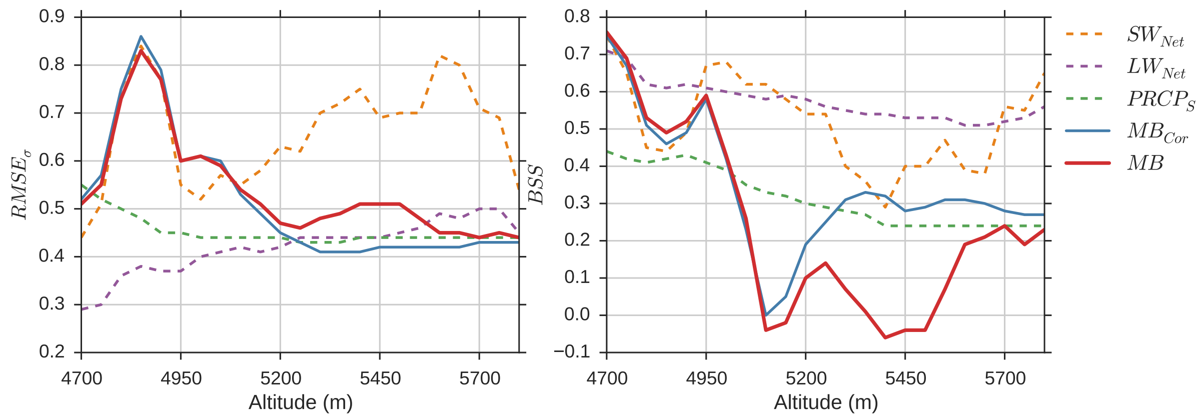

Figure R5: Updated Fig. 9, with a new diagnostic variable: MB_{Cor}, computed with a corrected melt energy flux conserving the mean during the calibration period.

As you correctly assumed, the introduction of this factor has a positive impact from the ELA at 4950 m a.s.l upwards, and it is particularly visible at high altitudes. However, it has a negative impact at low altitudes where it constrains the variability of the model. Altogether, introducing this correction factor has only a small impact on our conclusions (r^2 between glacier-wide specific MB and SST of 0.55, instead of 0.52 before). Since this melioration is not significant in comparison to the other uncertainties, we decided not to introduce this extra level of complexity. However, we updated the manuscript's Fig. 9 and the text in order to mention that this correction is available in DownGlacier.

4. P3010, lines 16-17: this statement depends on your classification (see general comment # 1). Actually, it is not true that Niños are often immediately followed by Niña (3 times over 7 years on Fig 2) and due to the classification applied here, Niño are always one-year long but it is usually not true: 91-93 is considered as a multi-year Niño, and some Niños are shorter than 1 year.

This statement was removed since we now follow a monthly classification.

5. P3011 line 15: how can the model better catch the precipitation inter-annual variability than the reference climatology?

The "reference" climatology is the reference model we chose to compare our downscaling with (Sect. 2.2.5). It is a simple average of all other months (i.e. the value for June 2007 is the average of June values in 2006, 2008 and 2009) and it is a coarse way to see if the downscaling model is better able to simulate inter-annual variability, i.e. able to predict that the next year will be wetter or not.

6. Fig 3: the agreement between full model and reference is good because the whole period serves as calibration. What are the results if only a two-year period is used as a calibration period, and the 2 remaining years as a validation period? And there are also some data from the AWS on the glacier (July 2010-Sept 2012) not used so far. I think it would be interesting to compare modelled data with observations during this period 2010-12 for validation.

The 2010–2012 period of measurements was affected by numerous data gaps and instrumental errors, especially for the radiation sensors. This prevented to run the SEB/SMB model for any longer than 2005–2009, and unfortunately also prevents any further validation of the downscaled fluxes. We agree that a four-year period for calibration / validation is short, but this is all we have currently...

We would like to emphasize that the full-model results you mention here are never used to assess the uncertainty of the model. All error bars are based on the cross-validation, which in this case is done with a "leave-5-out" algorithm (all 48 points of the cross-validation time-series are computed by 48 "penalized" models unaware of the 5 months surrounding the tested point). For indication, we provide several cross-validation time series in Fig. R6 below: from the simple leave-one-out to 2-fold cross-validation (4-fold cross-validation is synonym to having three years to calibrate and one to validate, and 2-fold cross-validation follows your suggestion of having two years for calibration and 2 years for validation). One can see that until the last case, the model is quite stable while with the 2-fold cross-validation, the model is largely unskilled. This can be explained by the climatic conditions of this four-year period: the two first and the two last years are very similar, thus strongly penalizing the model calibration.

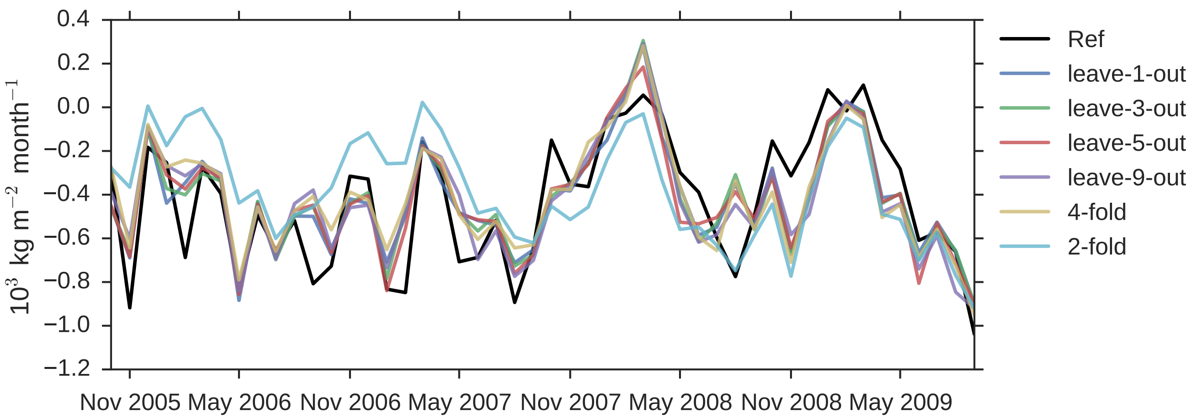

Figure R6: Same as the manuscript's Fig. 3 (bottom) but for various cross-validation methods.

From Fig. R7 below one can see that reducing the available period for calibration reduces the accuracy of the downscaling to a certain extent, but that the model is quite stable until the 4-fold cross-validation. The 2-fold cross-validation model is not unskilled for all variables, with the exception of precipitation and SW_{Net} *(not shown). Calibrating a 27 predictors model with 24 values is very demanding and these results are not surprising. We maintain the leave-5-out cross-validation scheme presented in the manuscript, as it appears to be a good compromise between a too optimistic estimation of the error (no cross-validation) and a too pessimistic one (2-fold cross-validation).*

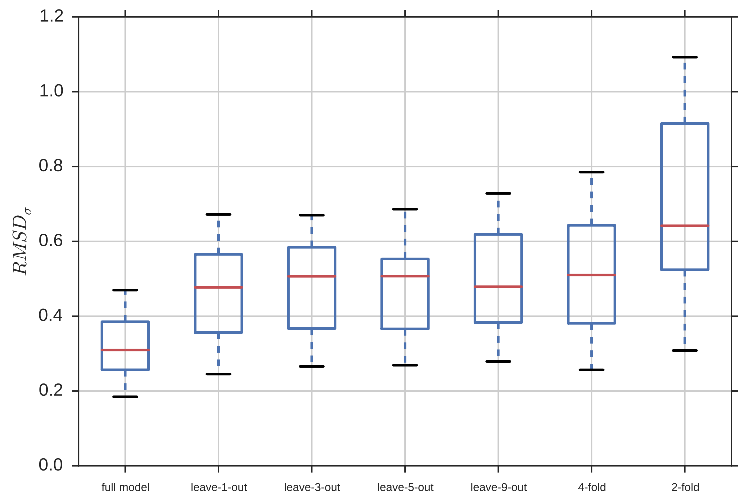

Figure R7: Box plots of $RMSE_\sigma$ for all downscaled variables and for various cross-validation methods.

7. P3012 line 5: how can modelled precipitation be negative? What does that mean?

This is due to the linear nature of the downscaling algorithm and is a known issue of all statistical *models (which are not aware of the physical properties of the variables). Since we are downscaling monthly values (which are close to a normal distribution), this does not happen very often (6% of the time for 1980–2013). When it happens, the predicted values are clipped to 0. We added this explicative sentence to the manuscript.*

8. Fig 4: it would have been interesting to see QS versus air temperature and QL versus wind speed as well.

The scatter plots are shown below and have been added to Fig. 4 in the manuscript. Since monthly air temperature is always close to 0° and to the surface temperature values, the sensible heat flux is more dependant on wind speed than on air temperature. The latent heat flux is well correlated with wind speed which is in turn well correlated with air humidity.

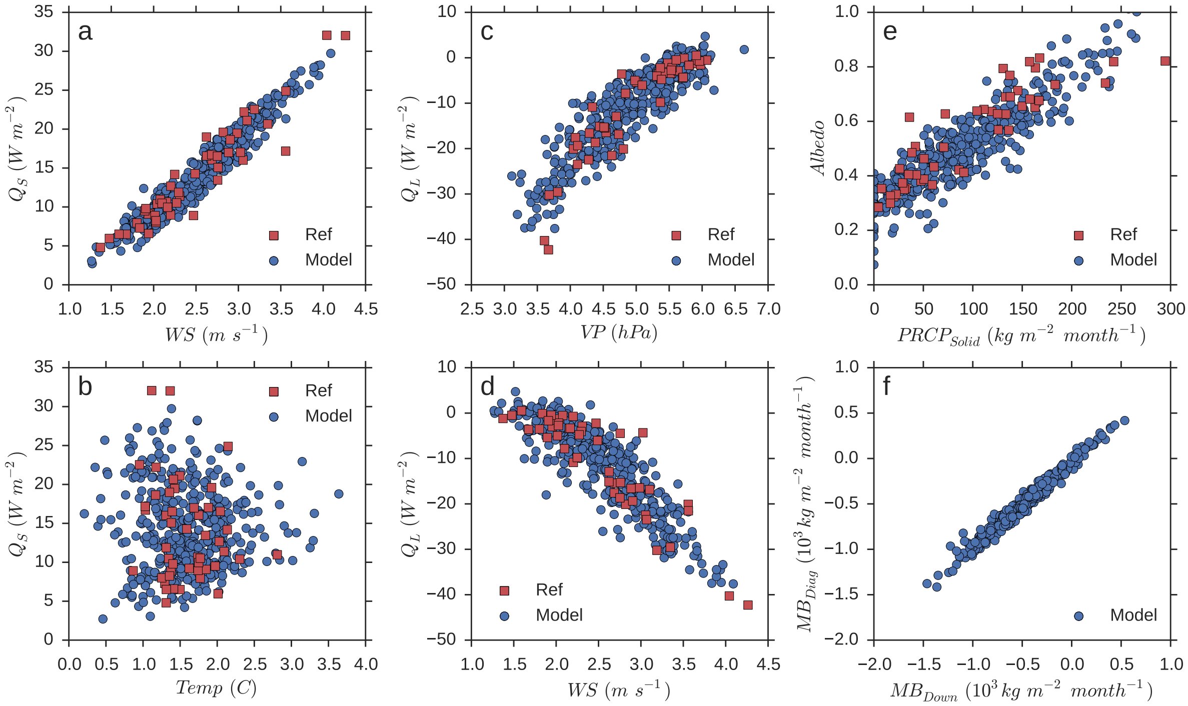

Figure R8: Updated Fig 04 with the new subplots: **(d)** sensible heat flux vs. air temperature, **(e)** latent heat flux vs. wind speed.

9. P3014, line 21. Francou et al (2004) based their analysis on Antizana Glacier (inner tropics) and obtained the best correlation between monthly MB and SSTA in Nino3.4 region applying a 3-month time lag. But for the outer tropics (Chacaltaya Glacier), Francou et al (2003) obtained the highest correlation between MB and SSTA in Nino1.2 region with a 2-month time lag (revised to 4-month lag by Rabatel et al (2013)). Since Shallap glacier is located in the outer tropics, I would have intuitively expected a better relationship between MB and SSTA in Nino1.2 region with a 2 (or 4)-month time lag instead of SSTA in Nino3.4 with a 3-month time lag. Could you comment on that? And what are the results with SSTA (Nino1.2; 2-month lag)? and with SSTA (Nino1.2; 4-month lag)?

The choice of the Niño 3.4 index appeared natural because it was used by Vuille et al. (2008b) in the Cordillera Blanca region. Furthermore, it is the indicator chosen by Trenberth (1997) *in his widely accepted "definition of El Niño". We agree that it is a bit arbitrary, and we now include the indexes you suggest. The three indexes are plotted in Fig. R9.*

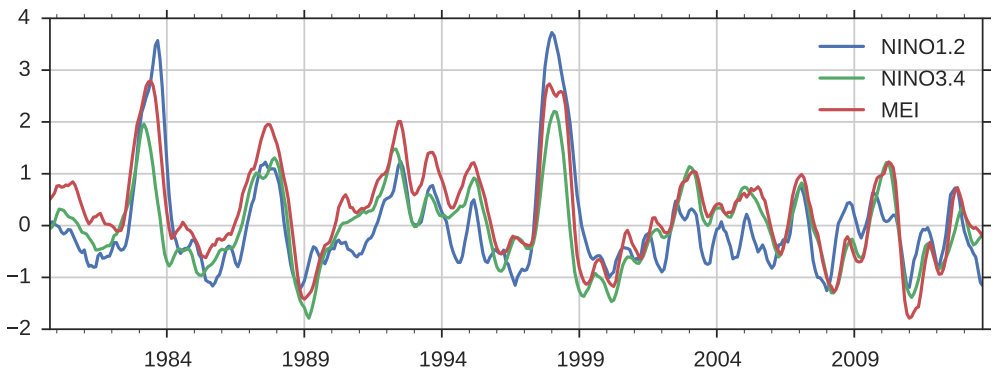

Figure R9: Time series of the Niño 1.2, Niño 3.4, and multivariate ENSO index (MEI).

The three indicators are close to each other, especially the MEI and Niño 3.4 which correlate with $r^2 = 0.88$ *on monthly values (Niño 1.2 and MEI correlate with* $r^2 = 0.61$*). Still, it is interesting to have a look at their predictive skill for annual MB at 4750 m a.s.l (see Table R1, added to the manuscript). It appears that Niño 3.4 and MEI both have very high correlations with MB, with a maximum at lag 2 and 3. Niño 1.2 has a lower predictive skill which maximises at lag 2. The differences between Niño 3.4 and MEI and between the lag values are far below the uncertainties of our computed MB, so that nothing very conclusive can be said about taking one or the other indicator. For simplicity we keep our initial Niño 3.4 classification but mention these new results in the updated manuscript.*

Table R1: Coefficient of determination (r^2) between the computed annual MB at 4750 m a.s.l and various ENSO indexes, for monthly lags between 0 and 5.

Lag	**0**	**1**	**2**	**3**	**4**	**5**
Niño 1+2	0.44	0.47	0.49	0.47	0.43	0.36
Niño 3.4	0.68	0.76	0.79	0.80	0.79	0.77
MEI	0.71	0.77	0.80	0.79	0.77	0.72

10. P3015 line 23-24 : What is the reason for negative MB at 5450m? Over-estimated melt (see specific comment # 3 above) or under-estimated solid precipitation, or both?

This is mostly related to the over-estimated melt. The correction factor introduced above (specific comment # 3) mitigates this problem.

11. P3016 line 1-13: it is a true problem while considering tropical glaciers where this transient snow cover in the vicinity of the ELA is responsible for a large variability of albedo, and consequently of the MB, especially during the wet season (which is the main melting and accumulation season and in turn the important season for the MB). According to me, this should be less a problem for mid-

latitude glaciers, because in summer, the surface state will be less variable in time. May be I misunderstood here, but I do not understand why the authors say that DownGlacier will perform poorly on mid-latitude glaciers (line 11). I believe that their conclusion i.e. to complement this approach with physical albedo models (line 13) is more justified for tropical glaciers. Since the model in unable to reproduce abrupt changes of MB from one month to the other, in the vicinity of the ELA, due to surface conditions, which is a key process over tropical glaciers controlling the MB, I do not understand why finally it performs rather well (line 15). I am wondering if there is any error compensation here, with for instance an over-estimation of modelled accumulation in the upper part of the glacier compensated by a too high melting in its lower part? It might be instructive to show the gradient of MB as a function of elevation. P3016, line 22: MB (1980-2013) = 0.04 ±0.4 m w.e./yr. It looks high (Rabatel et al (2013) give values closer to -0.6 m w.e./yr over this period see their fig 8) and difficult to explain only by the fact that glacier geometry is assumed unchanged. Any explanation for this?

The downscaling model will perform well in two situations:
- transient snow cover shorter than one month (always the case bellow 4800 m a.s.l.). i.e where there is a strong relationship between precipitation and albedo (Fig. 4 (a) in the manuscript)
- permanent snow cover
These conditions are met for a large part of Shallap glacier and are characteristic for tropical glaciers where temperature variability and annual cycles are low. In the mid-latitudes, there is a transient snow cover from winter to the beginning of summer, and a statistical model such as ours will not be able to simulate it. In the manuscript we mention the extreme example of a Tibetan glacier, where Mölg et al. (2014) showed that spring snowfall conditions are determinant for the entire annual mass-balance due to the albedo feedback.

We agree that this is a major source of uncertainty in our analyses, and this is why we concentrate most of our analyses (Figs. 3 to 8 and 12) at 4750 m a.s.l. and why we arbitrarily multiplied the glacier wide RMSD by a factor 2 for all glacier-wide analyses. The uncertainties are mostly present in the MB variability and not in the average (as nicely shown by the manuscripts' Fig. 10 (a), where the model searches for a compromise between over-estimation during the first two years and under-estimation afterwards, leading to a "correct" mean). The discrepancy between regional MB and our specific MB has been discussed in general comment # 2 above and is not related to this issue. We added a discussion paragraph in the revised manuscript in order to clarify these points.

Technical corrections

1. P3000; P3001, line 7 remove the right at the beginning of this line
2. P3003, line 14 and everywhere in the text and reference list : Francou et al., 2003

Done.

3. P3010, line 5 : the base period is 1980-2010 (caption fig 2) or 1981-2010 (line 5)?

1981–2010 (`http://www.cpc.ncep.noaa.gov/data/indices/` dataset ERSSTv4). Thanks.

4. Tab2: what is MSub? I believe it is MSubs i.e. subsurface melt (eq. 2)
5. Fig 13: y axis : add SSTA at the end of the legend

Done.