

Response to reviewers

Quantifying late-Holocene climate in the Ecuadorian Andes using a chironomid-based temperature inference model

We wish to thank everyone who contributed to the improvement of this manuscript, specifically the comprehensive suggestions from the two reviewers. Outlined below is a detailed response to both the reviewers' general, and specific comments. The manuscript has been substantially changed as a result of the reviewer's general comments (see below; *response to general comments*) and minor edits have been corrected and recorded accordingly. All typos and minor formatting errors, highlighted in the annotated pdf have also been changed.

Response to General comment: *reliability of the reconstruction.*

Whilst both reviewers, and an independent author who contributed to the online discussion, commented on the value of the study, all had major concerns relating to the final environmental reconstruction from Laguna Pindo and the subsequent interpretation. We would agree with all the reviewers that the development of a chironomid transfer function for the tropical Andes is an important contribution to tropical palaeolimnology and paleoclimatology. We also acknowledge, however, that chironomid studies from the tropics remain rare and little is known about the autecology of many of the taxa, namely their ecological tolerances relating to climatic variables. As a result, the environmental reconstruction from Laguna Pindo has some issues, namely unrealistically cold temperatures and significant inter sample variability. We would agree with all reviewers, and acknowledge in the original manuscript, that many of these fluctuations most likely relate to issues with the transfer function and/or the fossil record. These could include:

- The response of secondary variables, namely precipitation.
- Low head capsule concentrations in many of the samples.
- Un-even distribution of calibration lakes over the environmental gradient due to the steep topography of the Andes.
- Taxonomic issues (i.e different species between fossil samples and modern samples that currently cannot be separated using only larval head capsule material).

Many of these limitations are discussed in the manuscript. Indeed, we would argue a central point of this work would be the comparison of WA and Bayesian methods, in order to further explore these limitations. The application of the Bayesian model results in a less variable reconstruction, and an explanation for why the uncertainty associated with the reconstruction is greater than the climate variability we are reconstructing. The individual likelihood function of fossil taxa and the resulting posterior probability distribution for temperature sheds light on how the un-even distribution of calibration lakes, and subsequent skewed distribution of taxa, is affecting the inferred temperatures. Many taxa have unrealistically cold temperature optima due to the over representation of cold lakes in the calibration datasets and this has a significant affect on the reconstruction, most notably unrealistic cold Holocene temperatures. The error associated with both reconstructions is entirely consistent with a constant temperature of 20°C. We would agree with the reviewers that attributing the variability of the reconstruction to anything more than noise would be

an overstatement at this point. This work does, however, suggest the way forward for improving temperature reconstructions, namely, improving the richness/sampling of the training set to enable the detection of smaller signals.

For this reason we would agree with the anonymous reviewer; “ *These problems are honestly discussed in the text. It appears that L.Pindo was not the best lake to perform a reconstruction.*”. The reviewer provides two options for rectifying these issues:

“(i) The TF is optimized, undergoes additional testing, the quality of the reconstruction is substantially improved (robustness of the TT amplitudes, robustness of the cold anomalies, etc.) and/or (ii) the profile of the reconstruction is lowered; given the pertinent deficits the reconstruction is qualitative and NOT quantitative, not overstating the results and conclusions.”

Unfortunately, we do not feel that the quality of the reconstruction can be substantially improved at this stage for many of the reasons discussed previously and therefore we cannot meet first criteria (i). For this reason we propose to move forward with the reviewer's second suggestion, i.e. lowering the profile of the reconstruction in the manuscript. We feel this option will allow the manuscript to make a meaningful contribution to the literature, whilst honestly representing the current state of chironomid research in the area and addressing many of the concerns of all reviewers and online contributions.

The following major changes have been made to the manuscript in order to address the general comments of both reviewers:

- i) The Introduction has been shortened (L.126-L132) in order to reflect the new focus of the manuscript, i.e. refining the proxy as opposed to palaeoclimate inferences. This modification also addresses a concern of the anonymous review, which noted the introduction as being overly long.
- ii) Sub section 5.4 *Laguna Pindo temperature reconstructions* has been removed. This subsection is no longer needed as the temperature reconstruction is presented as qualitative and only used to further understand the various models.
- iii) Sections 5.6 (*Cooling climate 3800-2800 cal yrs BP*), and 5.7 (*Recent cooling*) have been removed. Based on the recommendations of the reviewers, we have changed the focus of the manuscript to center on proxy development not palaeoclimate interpretations. The Laguna Pindo reconstruction is used to understand the limitations of each model and is presented as a more qualitative interpretation of climate variability over this time period. The conclusions of the manuscript focus on our future recommendations for improving palaeotemperature inferences using chironomids. This addresses the current limitations of Neotropical palaeolimnology using chironomids, and provides a list of necessary criteria for future researchers wishing to explore this proxy further.

Response to specific comments

Line Number	Reviewers comment	Response
L. 38 R2	General Comment Due to the limitations of the environmental reconstruction the anonymous review suggest reference to the reconstruction should not be “quantitative”	“...the first quantitative reconstruction...” has been removed
L. 349 R1	I do not think that that lakes located between 1000 and 300 m asl can be considered to be similar in elevation to Laguna Pindo (~ 1200 m asl). Applying a standard lapse rate to this elevation range suggests that MAT for the lowest and highest lakes would vary by 12-20oC.	“...although the samples plot within the range of modern calibration lakes that lie at similar elevations (1000-3000 m a.s.l).” has been removed
L. 415 R1	Reporting the RMSEP as a % of the total MAT range captured by the training set would be useful.	Sentence becomes; “Although both models (WA inverse and Bayesian) perform well (WA RMSEP= 2.4°C/ 9.6% of training set range and Bayesian RMSEP= 2.3°C/9.2% of training set range)...”
Fig 1 R1	requires a N-arrow	N arrow has been added
Fig 2 R1	“PH” should be corrected.	PH has been corrected to pH
R1	It is not clear why non-limnological variables such as latitude were included in the exploratory analysis. Latitude, longitude and elevation are not directly controlling the distribution of midges; the analyses should be re-run with only environmental variables that have the potential to directly control the distribution of midges included.	These variables have been excluded from the analysis.
R2	The Introduction could be shortened (quite lengthy).	L.126-L132 have been removed from the introduction and the manuscript shortened to reflect the new direction of the paper.
R2	Chapter 3: I would not make too many sub-chapters (only one paragraph in 3.4 and 3.5)	Sub headings 3.4; 3.5; 3.6; and 3.7 have been removed

R2#	The sampling design for the downcore analysis should be described in detail (continuous, discrete sampling, regular intervals, stratigraphically. . .?) What is the percentage of sediment that is actually covered in the analysis? (e.g. 1 cm slice every 10 cm sediment makes 10% coverage and 90 % is not covered; this has serious implications regarding the robustness of the reconstruction).	Sentence added; <i>“The sampling interval for chironomid analysis was not uniform due to a varied sedimentation rate a varying sedimentation rate. To achieve as even a coverage possible over the time interval, samples were taken between every 10 and 20cm.”</i>
R2	I would include the Suppl Fig (Chronology) in the manuscript.	Table S1 and Figure S1 have now been included in the manuscript.
R2	Title: reconcile. It is quantitative indeed, but how robust and how good are the numbers? => Qualitative	Title has been changed to; Inferring late-Holocene climate in the Ecuadorian Andes using a chironomid-based temperature inference model
L.77 R2	Shulmeister	Corrected
L.93 R2	... preceding Glacial and Late-Glacial period ... (if you refer to 25-11.7 kBP; 20-25 kBP is not Late Glacial)	Changed to; <i>“(... c. 15,000-11,700 years before present...)”</i>
L.95 R2	Make also reference to Marcott et al 2013. This is the most comprehensive dataset.	Marcott et al 2013 has been added.
L.99-100 R2	Growing evidence from the tropics? I’m not sure about this. In fact it is still very controversial whether cold events (depending on the time scale) were globally, hemispherically or regionally synchronous (Wanner et al. 2011 QSR, Neukom et al 2014. NatCC;PAGES 2k 2013). The PAGES 2k Consortium 2013 has shown that, with a few exceptions (with solar-volcanic downturns) multi-decadal long cold phases were not coherent across the globe. Maybe rephrase sentence.	Changed to; <i>“Some evidence from the tropics suggests Holocene climate fluctuations such as the LIA are maybe global events....”</i>
L.108-109 R2	References not appropriate (these are not climatology papers). Make reference to Garreaud et al 2009 or Stefan Hastenrath 1991 Climate Dynamics of the Tropics or similar.	References have been modified.
L.151	How reliable are WT in a 10 cm deep water	In producing the manuscript we ran

R2	body? It should be assessed how sensitive the TF is with/without such lakes. In such water bodies the difference between MAT and WT is typically very large (in particular Tmax). I guess that the TF stats could be improved.	the transfer function using multiple combinations of different lakes included and excluded. This included removing very shallow lakes and overly deep lakes. The results presented are for the best performing inference model. We believe the problems which are leading to the unreliable reconstruction are overwhelmingly those discussed with reference to the all reviewers general comments. The manuscript has been changed accordingly to address this.
L.154 R2	. . . uppermost 1-2 cm . . . representing 5-20 years. . . Well, it was done like this and is usually done like this. But this implies that the sample for the Training Set depicts in one lake interannual/subdecadal variability (which may be very different from climatology!) and in another sample it is rather climatology (20-30 yrs). I suspect that this adds substantial errors to use the uppermost 3-4 cm of sediment to make sure that 20-30 yrs (climatology) are represented. The TT trends during 30 yrs are relatively small and similar in all lakes of the training set.	As pointed out by the reviewer the sampling method adopted here is common practice for chironomid studies of this kind. We would agree with the reviewer that testing the results of various sampling methods would be a worthy endeavor. The reviewer makes an important point that the uppermost sediments likely reflect inter-annual variability whilst deeper homogenous sampling may more accurately reflect climatology. Addressing this directly, however, would call for a complete re-sampling of the entire calibration dataset and will very probably not address the central concern of the reviewers; improving the reconstruction. This suggestion would not reduce the problems associated with un-even sampling.
L.171 R2	Fourteen 14C samples? Fig SOM shows six of them. Where are the others? Pls change and make it consistent with L 324 ff.	The Laguna Pindo record is much older and longer than the portion presented here. Much of the record is radiocarbon infinite and work is ongoing to produce a complete age depth model. Presented here is the portion of the record for which chironomid remains are found. This is addressed in Line 330-332 " <i>The best-fit age depth model for Laguna Pindo was a smooth spline (Fig S1). Due to the absence of chironomids at the bottom of the sequence, six radiocarbon samples were used for building the model with a total depth of the sediment considered of 461 cm</i> ".

		This figure has been removed from the SOM and placed in the manuscript itself.
L188 R2	Were nutrients (N and P) not measured? This might be a problem (Lotter et al 1998 J Paleolimnology)	Samples were taken for nutrients (anions and cation). Although filtered in the field, due to the remoteness of the fieldwork and continued biological activity, these samples were no longer reliable once returned for laboratory analysis in the UK.
L272 and 275 R2	avoid references in the results section. This reads like 'Discussion' L280 . . . optimum. . . (?)	References have been removed " <i>...optima</i> " changed to " <i>...optimum</i> "
L295 R2	in general, hc counts should be given in all Figures and Tables.	Hc counts has been added to all figures. The total number of head capsules for each calibration lake can be found in the data archive or <i>Matthews-bird et al 2015</i>
L295 f	is rather Discussion than Results. Move this paragraph.	Paragraph has been moved to discussion
L300 R2	Yes, this is critical (number of hc). It should be assessed whether the number of hc has an effect on the calibration statistics, in particular the residuals. (see also L303, I am not sure if this is the only criterion according to which the TF could be optimized)	We agree that the affect of head capsule concentration is extremely important, particularly with regard to WA models, which rely heavily on abundance. The Bayesian model, however, has a component of the model that uses only presence absence data. This was one reason for comparing the two methods. The Bayesian reconstruction and likelihood function, shows the effect of head capsule concentration on the reconstruction. Particularly the bias towards to colder temperatures. The current methodology already addresses the concern of the reviewer.
L.302 R2	Table 1 does not show these details (which are important), Table 1 shows the summary only. The details (hc) should be given (in the SOM)	Total number of head capsules for each lake can be found in the data repository.
L.327 R2	The sampling design must be clarified (in the Methods section). You took 30 samples spread over 420 cm. How did you take the samples? 1 cm slice every 10-15 cm? Stratigraphically (according to which criteria?) or continuously (complete sediment section)?	This has been addressed by a previous comment R2#.
L.353 R2	.. only seven samples? According to Fig 9 and the vertical dashed line there are many	This was a typo that has been rectified, 14 samples have a poor fit

	more.	to temperature.
L.485 R2	I don't think that anything is known about the precip/temperature relationship during the Late Holocene.	Sentence changed to; <i>"The location of Laguna Pindo makes it a good palaeoecological setting to record the response of temperature-sensitive proxies"</i>
L.495 R2	I think this is a substantial problem.	We agree with the reviewer that the lack of modern analogues is a substantial problem with the reconstruction. This is honestly discussed in the paragraph cited. This lack of modern analogues most likely reflects the uneven distribution of calibration lakes and the particular lack of lakes surrounding the fossil site.
L. 531 R2	according to this statement I would conclude that the temperature reconstruction of Laguna Pindo is qualitative at best.	We agree with the reviewer that more work is needed before Neotropical chironomids can be described as quantitative. As highlighted in our response to the general comments we accept that the profile of the reconstruction should be lowered. The passage now reads; <i>"The WA inverse MAT reconstruction, however, is statistically significant based on the criteria described by Telford and Birks (2011a) (Fig 10) suggesting that despite conflicting variables a temperature signal can be obtained from Neotropical chironomids although we caution against an over interpretation at this stage. Due to some of the limitations discussed previously, the reconstruction can currently only be deemed qualitative and requires more research."</i>
L. 539 R2	maybe also refer to Kanner et al (speleothems) and Ledru et al (N Ecuador)	This section has now been removed and significantly modified. The manuscript no longer over interprets the final reconstruction and these suggestions are no longer relevant.

L548 R2	Jones & Mann 2004 is not the best (has been criticized; S-Hemisphere is very poor). Suggestion: PAGES 2k 2013.	See previous comment
L.552/553 R2	I don't think this is true. There's a large body of literature pointing out the role of volcanoes, or a combination of S+V . . . rephrase sentence.	See previous comments
L.555 R2	No, I don't think this is true (cool from 400 yr BP onwards). The sample at 250 yr BP is still among the warmest of the entire record, almost as warm as today (!). There is only 1 sample (at 1850 AD) that shows cool conditions, and it is very questionable how robust that is (see your comment and my comments above)	See previous comment
L.569ff R2	It has been repeatedly demonstrated that the Andean ice cores (stable iso- topes) record precipitation and not temperature (as claimed by Thompson et al).	See previous comment
	LIA. Yes, this value has been reported for two Venezuelan glaciers (at 4600 and 5000 masl, mainly inferred from a drop in ELA by 300-500 m; Polissar et al. 2006). I doubt that similar (special high-elevation) conditions apply for L Pindo, given the limitations of the reconstruction (see above). This value seems extraordinarily high to me. Alternatively an explanation should be provided showing that such large TT amplitudes are physically plausible at local scales.	See previous comment
L.598 R2	Yes, the potential is shown (with the TF). But the reconstruction has major problems and severe limitations (see above). I would say: qualitative at best.	We agree with the reviewer and have modified the manuscript accordingly
L.605 R2	: . . .). Special. . .	Rectified
L624 R2	reference listed twice L634: Dryas-Holocene L667: check carefully L668: . . .Science 289	Duplicate reference removed
L.680. R2	Vol missing	Volume added
L.702 R2	Lemke	Rectified
L.810 R2	... Science 234, ...	Rectified

L.814 R2	Ref listed twice (also L819)	Duplicate reference removed
L. 818 R2	Holocene	Rectified
L.841 R2	Woodward, C	Rectified
L.855 R2	LOI: specify 550 or 950; ditto L858, Table 1 and Table 2, L883	550 has been sepcified
Table 1	Data set should be made available in full detail	Data is now available at Data Dryad
Table 3:	add units (where appropriate), also Caption Fig 5	Units added
Fig 2:	pH	PH changed to pH
Fig 3	(all Figs where appropriate, Fig 6, Fig 9): numbers of hc should be shown. It would be interesting to see the 'unusual lakes' (e.g. those with water depth of 10 cm).	Hc has been added to all necessary figures; fig 3,6,9