Reviewer 3
Main points

1.1 Scope
It seems that the goal of the paper is to evaluate the extent to which the potential nonstationarity of teleconnections for proxy-based reconstructions of ENSO. As such, there should be a more extensive review of such work. A non-exhaustive list would be: Stahle et al. [1998]; Braganza et al. [2009]; Wilson et al. [2010]; McGregor et al. [2010]; Emile-Geay et al. [2013a,b]; Li et al. [2011, 2013], few of which are acknowledged here.

- The goal of this paper is to evaluate the potential nonstationarity of teleconnections of ENSO from proxy reconstructions in regions where ENSO inferences have been made from single locations. We recommend that multi-dimensional information in the form of spatial patterns of change through time should be considered. We now included a more extensive discussion of such multi-proxy reconstructions, including several of the referenced studies above, in section 6.

It also would seem natural to pick (at least) one of the networks used above and see how vulnerable they are to the changes in teleconnections identified in the paper, on the context of pseudoproxy experiments [PPEs Smerdon, 2011]. One wouldn't have to use fancy reconstruction methods for this: an analysis of the signal-to-noise ratio in the network and how it changes from century to century would be all that is needed.

- This was not within the scope of our present study, which specifically aimed to investigate three regions using experiments with varying forcings. We would welcome the opportunity to investigate changes in teleconnection in pseudo-proxy experiments. We would gladly collaborate with the review on a study focusing on this suggestion.

On the topic of literature review, the authors should include more on volcanic effects on ENSO [Timmreck, 2012, and references therein].

- A detailed discussion of volcanics is now provided in section 6, including reference to Timmreck, 2012.

The section on ENSO characteristics (4.1) would do well to acknowledge the considerable work that has already been done to characterize ENSO in CMIP5/PMIP3 models. In particular Ault et al. [2013] showed that piControl simulations are incompatible with a suite of recent reconstructions [Emile-Geay et al., 2013a,b], while forced simulations are compatible, but seem to show a different phase relationship to the forcing. Also refer to Karnauskas et al. [2012] for a centennial-scale, ENSO-like oscillation that arises internally.

- We now include reference to Karnauskas et al 2012 and include discussion of previous work examining control and last millennium simulations in section 6.

1.2 Mechanisms
The main point of models is the ability to diagnose the causes of climate change. In this case, what makes teleconnections wobble, and is this robust across models? Do we expect the mechanism(s) to be stronger or weaker in nature?

- It is unclear what comparison is being suggested here. Are the models stronger or weaker in nature than *what/where*? In this study, the models are not used to diagnose the causes of climate change, but rather to investigate, as stated, whether proxy archives in the tropical Pacific are likely to be recording alterations in ENSO base frequencies or local-scale teleconnections under

differing boundary conditions. We now suggest that such diagnosis of mechanisms would make a useful future study ("Furthermore, our present study did not comprehensively investigate the relative influences on various external forcings (solar and volcanics) and internal variability on ENSO characteristic, which would provide useful information for comparison with proxy records. These mechanisms could be investigated, for example, using a suite of simulations with single or varying forcings.")

1.3 Statistical Considerations

Reference period It is good that the authors considered 100-year epochs within the past1000 ensemble, but it would have been logical to use a 100-year reference window for the historical or piControl simulations as well. I am surprised that they chose a 40-year span (1976-2006) and wonder how the results would change if they lengthened this reference period. For instance, the authors state "Although ENSO surface temperature anomalies across the Pacific are qualitatively similar, anomalies associated with the historical period (1976–2005) are generally of greater magnitude, particularly at remote locations outside the equatorial Pacific, including over North America and the south Pacific. These differences in magnitude between the Last Millennium and the historical may relate to the differing boundary conditions during the historical period associated with anthropogenic forcings, such as long-lived greenhouse gases, or simply from the greater diversity of ENSO episodes represented in the longer Last Millennium simulation." (emphasis mine). They need to rule out that this is not a sampling artifact due to comparing 100-year epochs to a 40y-long one.

- We have now included a 100 year historical period for analysis.

Statistical tests the Kolmogorov-Smirnov test is widely used to compare distributions, and I have no issue with its use here. I would only point out that the price of it not making distributional assumptions is that it has relatively low power. If the datasets are Gaussian, the authors may be better served by other tests that make this assumption, especially if they mainly intend to detect changes in location or scale. Note that precipitation is notoriously non-Gaussian, but can be made Gaussian via a transformation (cf the Standardized Precipitation Index, or SPI). Significance One of the most persistent problems in our field is that statistical tests are carried out assuming IID (independent and identically-distributed) data, which in many cases is not verified. Indeed, persistence from month to month or year to year often drastically reduces the number of degrees of freedom available for a test [Wilks, 2011]. Did the authors account for autocorrelation in tests presented in Fig. 4? Also, in Fig. 8, how significant are the variations in correlation? In many cases they look well within sampling error to me. It is imperative that the authors quantify this, because it is one of their main results ("it is evident in the model experiments that differing teleconnections may result at different points in time and may differ from present-day relationships"), and it may well evaporate in the face of statistical rigor. Do the correlations change sign altogether? How much would this bias a multiproxy reconstruction of ENSO?

- We did not account for autocorrelation in Figure 4. However, we have now included statistical significant in Figure 7/8 and accordingly in the text.

Wavelet spectra It should be noted that the Morlet wavelet spectrum as implemented by Torrence and Compo [1998] does not conserve energy, hence is not fit for spectral analysis [Liu et al., 2007]. The authors need to use the correction proposed in the latter paper and redo Fig. 2.

- On recommendation of all reviewers, we have reduced the length of the manuscript and focused on changes in the ENSO-local relationships through time. We have also removed the morlet wavelet spectrum plots.

1.4 Combining proxies

The idea to use multiple proxies to average out noise is nothing new. Few people will disagree with the authors when they write "We argue that proxy insights into change and variability in ENSO system are likely to be most robust when evidence is be synthesised over large spatial areas [...] considering multi-dimensional information in the form of spatial patterns of change through time is likely to yield more robust insights in large-scale systems."

While there could be many ways of synthesising evidence over large spatial areas, it seems that the authors have in mind the usual compositing, since they cite Li et al. [2013] as an example thereof. The authors should be aware that dating uncertainties may complicate this matter a great deal. Indeed, Comboul et al. [2014] showed that linear combinations of time-uncertain proxies may considerably distort the spectrum of the signal reconstructed from them. For certain diagnostics, like variance, McGregor et al. [2013] argued that one should first compute those diagnostics locally, prior to compositing. I ask that the authors acknowledge this work, and perhaps other efforts, to provide more specific guidance as to how one should synthesise evidence over large spatial areas in the real world. Such things are much trickier with real proxies than with gridded, exactly-dated GCM output.

- We do not suggest that multiple proxies should be averaged, and nor that we should "average out noise". We argue that using one location alone makes it difficult to determine where local, remote or teleconnected changes have occurred. The authors have considerable experience with generating palaeoclimate records and understand the limitations of dating material. We do not agree that the difficulties of generating proxy records means that interpretations that are not based of best practice are warranted. Furthermore, we have made specific recommendations that ENSO-related interpretations from remote sites "should be considered in conjunction with palaeo-reconstructions from within the central Pacific basin, the so-called "centre of action" of ENSO (Cobb et al., 2013)."

2 Editorial comments

The writing style is often long-winded. In many cases, this is because the authors are handwaving instead of basing their arguments on solid, quantitative proof. It is also rather imprecise and there are numerous omissions, some of which are pointed out here. A revised version should tighten up the writing.

- The specific recommendations detailed below have been included into the revised manuscript. Furthermore, we have attempted to make the writing more succinct. However, with few specific points of contention provided by the reviewer in terms of this comments, and the two other reviewers commenting that the manuscript is well written, these comments about handwaving and long-winded writing are very difficult to address explicitly.

1. "The MIROC-ESM model is excluded from this analysis as it exhibits large drift related error in the form of long-term trends that cannot be attributed to natural variability, but instead relate to deficiencies in model physics and numerics (Gupta et al., 2013) (Fig. 3)." (p1587, bottom) needlessly repeats p1584 L 29.

- This repetition has been removed from Section 3.

2. p1584 L26 "its representation of ENSO spectra is too short": improper terminology. Just say that ENSO is two biennial in this model, or that its dominant periodicity is too short. A spectrum is neither short nor long.

- This has been changed to read "One model (bcc-csm1-1) was excluded from analysis because its dominant ENSO periodicity is too short (Supplementary Fig. 4)."

3. p1588 L16-17 “compared with observed" change to "compared with observations"
- This has been changed.

4. p1589 L 16 “In the historical,..." . In the historical what?
- This has been corrected to say “historical experiment”

5. p1598 L 20 “a single climate model that well represents ENSO spatial dynamics, particularly on the western extent of the warm/cold tongue, would provide further insight into the apparent complexity of ENSO impacts through time.” Is there such a thing? I have yet to see a non-flux corrected CGCM whose Cold Tongue stays where it should be. Can the authors give an example?
- This has now been clarified to emphasise that sensitivity experiments with a single model would be useful, in addition to investigations using the CMIP5 ensemble (“First, additional targeted experiments within a single climate model would provide further insight into the apparent complexity of ENSO impacts through time.”)

6. The Bellenger et al reference has all author names duplicated : Bellenger, H., Bellenger, H., Guilyardi, E., Guilyardi, E., Leloup, J., Leloup, J., Lengaigne, M., Lengaigne, M., Vialard, J., and Vialard, J.: ENSO representation in climate models: from CMIP3 to CMIP5, Clim. Dynam., 42, 1999–2018, doi:10.1007/s00382-013-1783-z, 2013. Please check other references for similar mistakes.
- This has been corrected and references checked.