Reply to 'Review comments for hess-2018-173', Anonymous Referee #1

Comments: This paper provides very interesting results, and topic of the research is within the scope of this journal. The manuscript is well written and organized, but I have several minor comments as follows.
[General Comments] In the dynamic threshold method, three parameters (windows width: w, cut-off frequency: y and confidence level: alpha) should be specified, and sensitivity of w is discussed in this study. However, how about the sensitivity of y and alpha on extracting run length? # If authors could provide some comments/ideas on that, it would be helpful for readers. # (but do not have to conduct additional sensitivity analysis.)

Reply: We thank the reviewer for the comments and appreciate the opportunity to clarify aspects of the manuscript. Below we present our responses and indicate the changes made to the original manuscript.

The cut-off frequency parameter in the dynamic threshold method is used to filter out centennial trends that may be mixed with decadal variability. We therefore believe it is appropriate to use a cut-off frequency of 1/100 years. This parameter is aimed at filtering centennial trends and has little influence on the inferred runs which have significantly higher frequencies.

We have replaced Line 12 Page 10 in the original manuscript with "The cut-off frequency parameter in the dynamic threshold method is used to filter out centennial trends that may be mixed with decadal variability and should have little influence on the statistical characteristics of the inferred runs. A cut-off frequency of 1/100 years is considered to adequately meet these requirements."

The Mann-Whitney test is used to determine whether two independent sets of data come from the same distribution. There is a (100-α)% percent chance of the test statistic falling outside the α% confidence limits under the null hypothesis. If α is selected close to 100 and the test statistic falls outside the confidence limits, we are confident that the detected change point is actually a change point. The value of α is typically set by the researcher and involves consideration of type 1 and 2 errors (Zar, 1999). In this study, we follow the practice used in other reconstruction studies (Gedalof and Smith, 2001; Shen et al., 2006; McGregor et al., 2010; Pent et al., 2015) that adopt a confidence level of 90% so that the chance of making a type 1 error (that is, rejecting a changepoint when in fact there is one) is low, namely 10%.

We have added the following lines in the original manuscript at Line 14 Page 10 "In this study, we set the confidence level to be 90%, a value that is consistent with other reconstruction studies (eg Gedalof and Smith, 2001; Shen et al., 2006; McGregor et al., 2010; Pent et al., 2015)".

[Specific comments]
P13, P24-25: How did you set the value of "confidence level alpha (=90%)" and "cutting-off frequency (=1/11)" for this analysis? A little more explanation is expected (if possible). # P10, L14: y=100 (?)

Reply: See the above reply for our response on setting of the confidence level α.

The cut-off frequency of 1/11 is only used in Figure 7 to plot filtered PDV reconstructions. It is not used in the dynamic threshold method. The cut-off frequency of 1/11 was selected to be consistent with the cut-off frequency used in the Parker instrumental IPO time series.

P15, L13- "The absence of a consistent difference ... may be a consequence of sampling variability ... " –> The meaning of this part is a little unclear. # What is "sampling variability"?

Reply: Random samples from a given population are used to estimate target statistics for this population. Sampling variability refers to the variability in the target statistics that arises when random sampling is repeated. For example, two samples of 10 runs from the true run population will give different estimates of the run standard deviation. In this research, there are a limited number of PDV run samples. Therefore, it is important to recognize that differences in run statistics may be due to sampling variability. It is only when differences are bigger than what would be expected due to sampling variability that we would conclude there is a significant difference.

We have added the following text at Line 15 Page 15 to clarify this important issue:" Sampling variability refers to the variability in the target statistics that arises when random sampling is repeated. Therefore, it is important to recognize that differences in run statistics may be due to sampling variability. It is only when differences are bigger than what would be expected due to sampling variability that we would conclude there is a significant difference."

P28, Fig.1 What blue broken lines in (a4), (b4) and (c4) of Fig.1 are representing? #

95% confidence bands for zero autocorrelation (as described P7, L8)?

Reply: The dashed lines present the 95% confidence bands for zero autocorrelation. If the true autocorrelation were zero, there is 95% chance that the sample autocorrelation coefficient will lie between the dashed lines.

We have added the following text in Line 11 Page 28 Figure 1 caption: " The dashed lines in (a4)-(c4) present the 95% confidence bands for zero autocorrelation."

Reference:

Gedalof, Z., and Smith, D. J.: Interdecadal climate variability and regime-scale shifts in Pacific North 32 America, Geophysical Research Letters, 28, 1515-1518, 10.1029/2000GL011779, 2001.

McGregor, S., Timmermann, A., and Timm, O.: A unified proxy for ENSO and PDO variability since 20 1650, Climate of the Past, 6, 1-17, 10.5194/cp-6-1-2010, 2010.

Peng, Y., Shen, C., Cheng, H., and Xu, Y.: (2015): Simulation of the Interdecadal Pacific Oscillation and its impacts on the climate over eastern China during the last millennium. J. Geophys. Res. Atmos., 120, 7573–7585. doi: 10.1002/2015JD023104.

Shen, C., Wang, W. C., Gong, W., and Hao, Z.: A Pacific Decadal Oscillation record since 1470 AD 46 reconstructed from proxy data of summer rainfall over eastern China, Geophysical Research Letters, 47 33, L03702, 10.1029/2005GL024804, 2006.

Zar, J. H. (1999). Biostatistical analysis, 663 pp: Prentice Hall, Englewood Cliffs.

Reply to 'Review comments for hess-2018-173', Anonymous Referee #2

General comments

This is an interesting and mostly well-written manuscript with a topic of interest to HESS. My main comments concerns the presentation, which I think can be strengthened in places; in particular as the authors propose a new methods. On page 11, lines 20-23 it is stated that Vanc15 is more reliable than Macd05 and Mann09. As such, does the proposed threshold method help to cover-up problems with the relatively poor quality of some of the reconstructions? If one of the reconstructions is clearly more reliable than others, should the advice not be to use Vanc15 in future studies and not Macd05 or Mann09?

Reply: We thank the reviewer for the comments which have helped improve the manuscript and appreciate the opportunity to clarify aspects of the manuscript. Below we present our responses and indicate the changes made to the original manuscript.

The threshold method does not hide problems with poor quality reconstructions. Indeed Figure 7 clearly shows the reconstructions produce quite different run time series. However, reconciling these reconstructions was not the objective of the paper. Rather the objective here is to develop a robust run length extraction method to maximize the potential value of the reconstructions and to investigate the consistency of the statistical signatures with regard to two hydrologically important questions,: (i) are PDV runs lengths stationary and (ii) is the persistence (i.e. run lengths) of wet and dry epochs different.

In the last paragraph of the conclusions, we comment that the reconstructions provide quite different PDV run length time series and that reconstructing the palaeo PDV run length time series from multiple reconstructions will need careful consideration of errors. For example, the approach presented by Henley et al. (2011) combines individual reconstructions according to their accuracy. However, the accuracies are not high with Nash-Sutcliffe indeces < 0.48, suggesting even favouring the more accurate individual reconstructions still results in a combined/overall reconstruction with large errors. So our approach is to look for a signal that is consistent in all the individual reconstructions.

We have added the following text after Line 23 Page 11 in the original manuscript: "Individual reconstructions are subject to different error structures and magnitudes. Use of multiple reconstructions can help reduce the impact of errors, and thus provide more insight into the statistical characteristics of PDV run lengths. Nonetheless, errors still need to be carefully considered. This is illustrated by Henley et al. (2011) who combines individual reconstructions according to their accuracy. However, the accuracies of the individual reconstructions are not high (Nash-Sutcliffe indices <0.48) suggesting that even favouring the more accurate

individual reconstructions will still result in a combined/overall reconstruction with large errors. Therefore, the focus of our research is to identify a signal (or signals) that is (are) common to all individual reconstructions."

On Line 18 Page 4 we cited von Storch et al. (2004, 2006) and Wahl et al. (2006) who considered the problem of whether to use detrended or nondetrended data. We have added the following sentences on Line 18 Page 4 to expand the discussion of von Storch et al. (2004, 2006) and Wahl et al. (2006). "As stated by von Storch et al. (2006), "it is commonly accepted that proxy indicators may contain nonclimatic trends". Moreover, the calibration and validation of any statistical method using nondetrended data may be compromised because the nonclimatic trends may be interpreted as a climate signal. The centennial trends in PDV reconstructions may be either nonclimatic trends or non-decadal climate trends. Whichever the case, it is necessary to filter such centennial trends before interpreting decadal climate variability."

Is there a case for also looking at the magnitude of sojourns above and below a threshold in addition to duration?

Reply: We agree that looking at the magnitude of sojourns above/below a threshold would be of considerable interest. However, that is not the focus here as we are aiming to utilize PDV run length information to parameterise stochastic models used in water resources management related decision making.

Other comments

Page 2, line 14-15: semantics, but do you mean the impact of PDV has been explored at locations around the world?

Reply: Yes the impact of PDV has been explored at various locations around the world. As the impact of PDV is discussed in the paragraph starting Line 3 Page 3, we have deleted "and explored" in Line 13 Page 2.

Page 3, lines 8-10: Would be useful to see geographical extend and topic (flood/drought) for each of these references.

Reply: As the focus of the paper is not the PDV impact, which is now established, we feel the paragraphs starting at Line 12 Page 2 and Line 3 Page 3 provide sufficient information about the geographic spread of PDV and its impact context.

Page 4, line 11: what is the unit for the +-0.5 threshold? See also page 8, line 12.

Reply: The threshold is unitless as the PDV indices are standardized. To make this explicit, in Line 11 Page 4 "Vance et al. (2015) used ±0.5 as thresholds" is replaced with "After standardizing the indices, Vance et al. (2015) used ±0.5 as thresholds"

Page 6, line 21: SST not defined.

Reply: Agreed.
We replaced Line 21 Page 6 "monthly SST anomalies" with "monthly Sea Surface Temperature (SST) anomalies"

Page 7, line 21: define paleo proxies.

Reply: The original text in Line 21 Page 7 "The reconstructions are based on paleo proxies" is replaced with "The reconstructions are based on paleo proxies (i.e. preserved physical characteristics of the environment that can be directly measured)".

Page 8, lines 8-15: This paragraph is a very similar to page 4. Is it necessary to write all this twice?

Reply: Agreed.

Changes to original manuscript: Page 8, Lines 8-15 are deleted and Page 8 Line 16-17 are changed to "For reconstructions Macd05, Mann09 and Vanc15 which exhibit centennial trends (see Table 1 and Figure 2), the

above mentioned static threshold methods used in Verdon and Franks (2006), Henley et al. (2011), Vance et al. (2015) and Henley et al. (2017) may not identify meaningful decadal phases."

Page 8, lines 22-23: As per my comment in the introduction: is the use of a more dynamic threshold method simply masking underlying problems with some of the reconstructed datasets. If so, should these datasets not be excluded from further analysis on the basis that more reliable datasets are now available?

Reply: The dynamic threshold method only filters out centennial trend and has little influence on the inferred run lengths which have significantly higher frequencies. Please refer to the reply to the first comment for further commentary on this.

Page 9: I am not sure I understand the method to a level where I could implement my own version of the dynamic threshold framework; in particular, step 1. Maybe a conceptual figure assisting the reader could be useful here?

Reply: We explored a number of options but felt our efforts would not value add beyond the existing four-step description and Figure 2 which shows the PDV time series and phases determined using the dynamic threshold method.

Instead we have opted to provide the R code for the dynamic threshold method in supplementary information. Furthermore we have enhanced the description of step 1 as follows:
"1. Detect step-change points. For a given reconstruction, apply a change point detection method. A number of methods can be used. In this study we used the non-parametric Mann-Whitney test method (Mauget, 2003) with a given window width w and confidence level α to identify the step-change points. The method involves centring a window of width w at a particular year t and then applying the Mann-Whitney test to the samples of width w/2 at or before and after year t. A step change is deemed to occur if the Mann-Whitney test statistic lies outside the α confidence limits (under the null hypothesis)."

In addition, we have referred to Figure 2 earlier. In Line 3 Page 9 we have added: "Figure 2 provides insight about the mechanics of dynamic threshold method showing the PDV time series and the resulting block phase waveform."

Page 9: I have not previously come across a Butterworth filter. Are Eqs 1 and 2 filters of this type?

Reply: Eq 1 is the arithmetic mean of the PDV index for a particular run. It does not use the Butterworth filter. Likewise Eq 2 is an arithmetic mean. However, in this case, it is taking the average of the Butterworth filtered index values. We chose the Butterworth filter to maintain consistency with Henley et al. (2011) who selected the filter on account of its near-flat frequency response in the pass band. In Line 15 Page 9 we replace "(Selesnick and Burrus, 1998)" with "(which was used by Henley et al. (2011) to filter paleo PDV indices)".

Page 10, line 4: unit on y=100? Section 3.2.2: Last sentence (lines 11-13) is rather meaningless and not necessary I think.

Reply: In Page 9 Line 17 it is stated that "cut-off frequency 1/y year^{-1}", in which y is a unitless parameter.
Changes to original manuscript: Last sentence (lines 11-13) in Section 3.2.2 is deleted.

Page 13: I am not clear on what is the difference between the conditional and unconditional distributions.

Reply: We have removed reference to "unconditional" but have retained usage of conditional. A conditional distribution is the distribution of runs that meet a specific condition. For example, a distribution conditioned on runs less than 20 years would exclude all runs greater than 20 years.

Changes to original manuscript: Page 13 Line 8-11 are changed to "Panel (a1) shows the empirical density of run lengths, while panels (a2) and (a3) show the empirical densities conditioned on run lengths less than 20 years and greater than 20 years respectively. Likewise Panels (b1) to (b3) repeat this sequence for 20 and 60-year windows with the conditioning threshold of 30 years".

Figure 6: There is a lot of information on this figure but the overall summary is summarised nicely on page 17-19. Could this figure somehow be simplified to be more focussed on this message?

Reply: We thank the referee for the positive comment on our summary. However, we cannot see any way to further simplify Figure 6 - removing any feature from the plot would leave some of our conclusions unsubstantiated. The plots highlight the critical influence of window width on the inferred run length distributions.

Page 13, line 24: Is alpha not a significance level (\alpha=10%) rather than a confidence level as it refers back to a statistical test?

Reply: We accept that our usage of "level" may cause confusion. We have replaced "confidence level" with "confidence limits". As explained in an earlier response, if the test statistic falls outside its α confidence limits (under the null hypothesis), a step change is deemed to have occurred.

Page 13, line 25: On page 10, line 14 y is defined as 100, but here y=11; why the difference?

Reply: The cut-off frequency of 1/11 is only used in Figure 7 to plot filtered PDV reconstructions. It is different to the parameter y in the dynamic threshold method, which is used for filter out centennial trend. The cut-off frequency of 1/11 was selected to be consistent with the cut-off frequency used in the Parker instrumental IPO time series.

Section 4: Not sure the title of this section is appropriate as 'hydrology' is not discussed at any point.

Reply: In the introduction we stated that positive/negative PDV phases are related to dry/wet hydrological epochs. This association motivated the original title. Nonetheless, we accept the reviewer's point and have changed the section title to "Statistical signatures of PDV run lengths".

Page 15, line 9: What does 'pooled' mean? Is this run-lengths extracted from all 12 reconstructions merged together into one sample?

Reply: Yes, pooling does mean that all 12 reconstructions are merged into a single sample set. Please note pooled positive runs refer to merging all the positive runs in the 12 reconstructions into a single sample set.

Page 15-16: There is a lot of information on Figure 9, but I am not sure I understand how to interpret them. Also, are the conclusions derived from Figure 9 really that different from what you have already found in the much more easily interpreted Figure 8, that there is little evidence of statistical significant differences? Same comment on the difference between Figures 10 and 11.

Reply: Sampling variability makes interpreting differences particularly challenging when working with small samples, the danger is to avoid attributing significance to what is likely to be noise. For this reason, we feel that analysis of differences from different perspectives helps guard against such an event. The fact that the conclusions drawn from Figure 9 are in line with Figure 8 (same with Figures 10 and 11) provides more confidence. For that reason, we believe Figures 9 and 11 offer value and should be retained.

Page 15: Symbols mu+, mu-, sigma+ and sigma- are not defined anywhere?

Reply: We changed Page 15 Line 22 into "Figure 9 presents the joint posterior distribution of the differences in mean (μ^{+}-μ^{-}) and standard deviation (σ^{+}–σ^{-}) for each reconstruction"

Page 18, lines 4-5. Not sure I understand the meaning of this sentence.

Reply: We have replaced the first sentence is section 4.3 with "All of the reconstructions reported in Table 1 did not distinguish between positive and negative PDV runs in their analysis, except for Vance et al. (2015)."

Reference:

von Storch, H., E. Zorita, J. M. Jones, F. Gonzalez-Rouco, and S. F. B. Tett, 2006: Response to Comment on ‘‘Reconstructing Past Climate from Noisy Data". Science, 312, 529, doi: 10.1126/science.1121571.

Using paleoclimate reconstructions to analyse hydrological epochs associated with Pacific Decadal Variability

LANYING ZHANG, GEORGE KUCZERA

School of Engineering, University of Newcastle, Callaghan, New South Wales, Australia

ANTHONY S. KIEM

Centre for Water, Climate and Land (CWCL), Faculty of Science, University of Newcastle, Callaghan, New South Wales, Australia

GARRY WILLGOOSE

School of Engineering, University of Newcastle, Callaghan, New South Wales, Australia

Abstract

The duration of dry or wet hydrological epochs (run lengths) associated with positive or negative Inter-decadal Pacific Oscillation (IPO) or Pacific Decadal Oscillation (PDO) phases, termed Pacific Decadal Variability (PDV), is an essential statistical property for understanding, assessing and managing hydroclimatic risk. Numerous IPO and PDO paleoclimate reconstructions provide a valuable opportunity to study the statistical signatures of PDV, including run lengths. However, disparities exist between these reconstructions making it problematic to determine which reconstruction(s) to use to investigate pre-instrumental PDV and run length. Variability and persistence on centennial scales are also present in some millennium long reconstructions, making consistent run length extraction difficult. Thus, a robust method to extract meaningful and consistent

run lengths from multiple reconstructions is required. In this study, a dynamic threshold framework to account for centennial trends in PDV reconstructions is proposed. The dynamic threshold framework is shown to extract meaningful run length information from multiple reconstructions. Two hydrologically important aspects of the statistical signatures associated with the PDV are explored: (i) whether persistence (i.e. run lengths) during positive epochs is different to persistence during negative epochs and (ii) whether the reconstructed run lengths are stationary during the past millennium. Results suggest that there is no significant difference between run lengths in positive and negative phases of PDV and that it is more likely than not that the PDV run length has been non-stationary in the past millennium. This raises concerns about whether variability seen in the instrumental record (the last ~100 years), or even in the shorter 300-400 year paleoclimate reconstructions, is representative of the full range of variability.

Key words: PDV; run length; multiple reconstructions; stationarity

1. Introduction

A pattern of Pacific ocean-atmosphere climate variability at decadal time scales has been identified and explored at various locations around the world including Africa (Reason and Rouault, 2002;Hoell and Funk, 2014), Eastern and Southern Asia (Krishnan and Sugi, 2003;Ma, 2007), America (Mantua and Hare, 2002;Andreoli and Kayano, 2005) and Australia (Kiem et al., 2003;Verdon et al., 2004;Verdon and Franks, 2006;Henley et al., 2011). This pattern, referred to in this paper as Pacific Decadal Variability (PDV), is associated with sea surface temperature fluctuations and sea level pressure changes in the north and south Pacific Ocean.

PDV is usually described by the Inter-decadal Pacific Oscillation (IPO) or Pacific Decadal Oscillation (PDO) indices (Mantua et al., 1997;Power et al., 1999). Of particular hydrological relevance are the statistical characteristics of PDV phases where a phase refers to a period during which the PDV index lies above (or below) some thresholds (Dong and Dai, 2015;Verdon et al., 2004;Henley et al., 2011;Vance et al., 2015). The duration of a PDV phase (termed run length) is

defined as the time between consecutive crossings of the threshold. The phase may be described as positive (negative) if the PDV index is above (below) the threshold, or dry (wet) if the PDV phases are associated with predominantly dry (wet) hydrological conditions.

Of particular relevance to the hydrology and water resources community is the evidence that the PDV phases can be associated with multi-decadal periods of persistently wetter or drier conditions and corresponding increases in flood or drought risk in affected regions, particularly those on the Pacific rim. PDV has been found to be related to a number of hydrological variables including precipitation, streamflow, flood/drought risk (Cook et al., 2013;Kiem et al., 2003;Verdon et al., 2004;Dai, 2013;Goodrich and Walker, 2011;Hu and Huang, 2009;Li et al., 2012;McCabe et al., 2012;Mehta et al., 2011;Wang et al., 2014;Henley et al., 2013). For example, Kiem and Franks (2004) showed in a case study for eastern Australia that the probability of reservoir storages falling below a critical level differs significantly depending on the PDV phases (see Figure 6 in Kiem and Franks (2004)). Kiem and Franks (2003) and Micevski et al. (2006) demonstrated that flood risk in eastern Australia is strongly dependent on PDV phase. Henley et al. (2013) extended this work to show that short-term drought risk is strongly dependent not only on the PDV phase but also on the time spent in a particular phase.

Despite the clear relevance of PDV phases to hydrological risk assessment and water resource management, there remain considerable knowledge gaps about the statistical characteristics of PDV, including run lengths. The instrumental record shows that PDV phases have varied irregularly with runs ranging from less than a decade to several decades during the past century. However, the instrumental record is insufficient to characterize the statistical characteristics of PDV run lengths. In response to this, significant advances have been made in reconstructing pre-instrumental PDV behaviour to extend data length. For example, at least 12 IPO or PDO reconstructions have been published (Biondi et al., 2001;Gedalof and Smith, 2001;D'Arrigo et al., 2001;MacDonald and Case, 2005;D'Arrigo and Wilson, 2006;Shen et al., 2006;Verdon and Franks, 2006;Linsley et al., 2008;Mann et al., 2009;McGregor et al., 2010;Henley et al., 2011;Vance et al., 2015).

In view of these published PDV reconstructions, the fundamental question is whether useful information can be extracted about the statistical characteristics of PDV persistence. In dissecting this question, several unresolved issues are identified:

1) Static threshold methods have been used to estimate run lengths of PDV phases. This raises the concern that biased conclusions about the statistical characteristics of decadal climate variability may be drawn when reconstructions exhibit centennial or longer trends. The non-parametric Mann-Whitney test method was used in Verdon and Franks (2006) whereby a crossing was defined when the test detected a statistically significant difference between two halves of data in a 30-year moving window, with zero used as a static threshold to define the sign of PDV phases (Verdon-Kidd, personal communication, 24 April, 2017). Henley et al. (2011) used the instrumental mean of the composite IPO/PDO index~~, while~~ After standardizing the indices, Vance et al. (2015) used ±0.5 as thresholds to determine crossings and the sign of PDV phases, and Henley et al. (2017) used the long-term modelled IPO Tripole Index (TPI) mean as the static threshold with run lengths less than 5 years omitted. However, when static threshold methods are used, extraordinary long centennial run lengths may be identified in the reconstructions that exhibit centennial or longer trends (MacDonald and Case, 2005;Mann et al., 2009). Analysis based on data with long-term trends can lead to results that are overwhelmed by such trends and hence efforts have been made to detrend data without losing useful signals (Wahl et al., 2006;von Storch et al., 2006;Wu et al., 2007). As stated by von Storch et al. (2006), "it is commonly accepted that proxy indicators may contain nonclimatic trends". Moreover, the calibration and validation of any statistical method using nondetrended data may be compromised because the nonclimatic trends may be interpreted as a climate signal. The centennial trends in PDV reconstructions may be either nonclimatic trends or non-decadal climate trends. Whichever the case, it is necessary to filter such centennial trends before interpreting decadal climate variability. Given that static threshold methods cannot remove trends, how should the run length extraction method be designed to extract useful and consistent information from all reconstructions, including reconstructions exhibiting centennial trends? To what extent are run length distributions sensitive to the choice of extraction parameters (e.g. threshold, window width etc.)?

2) The durations of positive and negative PDV phases have not been treated separately in most previous research (e.g. Verdon and Franks, 2006; Linsley et al., 2008; Henley et al., 2011a) Using a high-resolution millennial-length IPO reconstruction, Vance et al. (2015) observed that positive phases dominate and last longer than negative phases. Owing to nonlinear hydrologic feedback mechanisms, such as elasticity of streamflow to runoff (Chiew, 2006) and shifted rainfall-runoff relationships (Saft et al., 2015), the duration of increased (decreased) precipitation during a wet (or dry) PDV phase may impact on streamflow in a highly nonlinear manner. Therefore, assuming wet and dry runs follow the same distribution may misrepresent the intensity of impacts. Is there sufficient evidence about the dissymmetry between positive and negative phase run lengths from the multiple reconstructions? Will run length samples from each phase lead to significantly different run length simulations?

3) Stationarity of climate signatures is a necessary, yet implicit, assumption underlying many analyses based on paleoclimate reconstructions – and hydroclimate stochastic modelling used in water resource assessment (e.g. Henley et al., 2013). However, a number of studies have suggested the possibility of climatic non-stationarity over the past millennium. Phipps et al. (2013) found that the relationships between paleoclimate proxies and climatic variables have been changing at least the last millennium. Razavi et al. (2015) explored the stationarity of the mean, variance and autocorrelation structures of paleo tree-ring proxy data and concluded that the key statistical characteristics of climate had undergone significant change over time. The existence of extraordinary warm and cold periods, known as Medieval Climate Anomaly (~950-1250CE) and Little Ice Age (~1400-1700CE) (Mann et al., 2009;Phipps et al., 2013;Atwood et al., 2015), also challenges the assumption of climatic stationarity. Furthermore, externally forced anthropogenic climate change, arising from elevated atmospheric concentrations of greenhouse gases, may also make recent and future climate different from the long past (Kirtman et al., 2013). Therefore, care should be taken when positing that PDV characteristics are stationary. Statistical signatures of past reconstructed PDV runs should be tested for non-stationarity (and where possible non-stationarity should be attributed to causal mechanisms) to ensure robust and representative estimation of the full range of variability.

In view of the above mentioned issues, this study has the following objectives:

1) To develop a more robust run length extraction method that is applicable to all reconstructions, including those with centennial trends. Of particular importance is the sensitivity of the inferred run length distributions to the choice of the extraction method parameters.

2) To identify hydrologically important statistical properties of PDV that are common to different reconstructions. In particular two fundamental questions are explored: (i) whether persistence (i.e. run lengths) during positive epochs is different to persistence during negative epochs and (ii) whether the reconstructed run lengths are stationary during the past millennium. Both questions address fundamental concerns about using PDV and run length information in stochastic hydroclimatic models used to assess drought and flood risk.

The rest of the paper is organized as follows. Section 2 briefly introduces the instrumental and reconstructed PDV records used in this research. The dynamic threshold run length extraction method is introduced in Section 3, along with its comparison with the static threshold method and determination of its parameters. The hydrologically important statistical signatures of PDV are analysed and discussed in Section 4, with conclusions presented in Section 5.

2. Data

2.1 Instrumental PDV records

A number of instrumental IPO and PDO indices have been published to characterize PDV since ~1900. There is a strong correlation between IPO and PDO indices, suggesting that both indices represent a similar broad pattern of climate variability. The primary difference between the IPO and PDO indices is that the IPO index is based on a broader spatial scale than the PDO index (Folland et al., 2002;Verdon and Franks, 2006;Henley et al., 2011).

Three PDV indices are used in this study: (1) The unfiltered instrumental PDO index (http://research.jisao.washington.edu/pdo/PDO.latest.txt), denoted as PDO_Mantua, is the monthly standardized value of the leading principal component of monthly Sea Surface Temperature (SST) anomalies in the North Pacific Ocean, poleward of 20N (Mantua et al., 1997;Zhang et al., 1997); (2) The unfiltered monthly IPO index from Parker et al. (2007), denoted as IPO_Parker, is the second covariance empirical orthogonal function of low-pass-filtered SST; (3) The unfiltered monthly Tripole Index (TPI) for the IPO (Henley et al., 2015), denoted as IPO_Henley, is based on the SST anomalies in three large geographic regions of the Pacific. Annual (calendar year) values of these three indices are taken as the average of monthly values.

Figure 1 shows time series of the three indices along with probability density plot, quantile-quantile (QQ) plot and autocorrelation plots. The distribution density plot and QQ plot of the annual PDV indices suggest that the instrumental PDV indices are approximately normally distributed with zero mean and unit standard deviation (except for IPO_Parker). The lag-one autocorrelation lies between about 0.3 and 0.5. By lag three, the autocorrelations lie within the 95% confidence bands for zero autocorrelation.

The instrumental PDV indices are used for two purposes in this study:

(a) to assess the similarity between IPO and PDO indices. If instrumental IPO and PDO indices can be used to represent the same broad pattern of climate variability - and noting that different reconstructions are calibrated against different instrumental indices - there is no need to recalibrate all reconstructions against the same instrumental data.

(b) to guide the selection of the parameters of the run length extraction method. If selected parameters can identify credible run lengths in the instrumental PDV record, it is assumed that they can also produce credible run lengths in the reconstructed record.

2.2 PDV reconstructions

Twelve published annual PDV reconstructions from different sites with different proxies are used in this study. Of these, nine reconstructed the last 300-400 years and three reconstructed the last ~1000 years. The reconstructions are based on paleo proxies (i.e. preserved physical characteristics of the environment that can be directly measured) from the northern and southern, and western and eastern regions of the Pacific Ocean. A summary of these reconstructions is presented in Table 1, while Figure 7 in Section 3.2.3 presents time series plots of the 12 reconstructed indices, and they are denoted throughout by the four-character author name and two-character publishing year. It is worth noting that Mann09 was not directly calibrated to an instrumental PDV index, and also has a negative correlation with other reconstructions. However, because the correlation of this reconstruction with the other ones is relatively high (Henley, 2017), the Mann09 reconstruction is retained with the sign reversed.

3. Run length extraction

3.1 Static and dynamic threshold run length extraction methods

~~A number of studies have attempted to extract phases (or run lengths) from PDV reconstructions. Verdon and Franks (2006) used the Mann-Whitney test method to determine whether a step change occurred within a 30-year moving window. Each step change point was treated as a potential crossing. They used zero as a static threshold to define the sign of PDV phases (Verdon-Kidd, personal communication, 24 April, 2017). The instrumental mean and ±0.5 were used as static thresholds in studies by Henley et al. (2011) and Vance et al. (2015) respectively, and later, Henley et al. (2017) used the long-term IPO Tripole Index (TPI) mean as the threshold to determine crossings and the sign of PDV phases.~~

However, for reconstructions Macd05, Mann09 and Vanc15 which exhibit centennial trends (see Table 1 and Figure 2), the above mentioned static threshold methods used in Verdon and Franks (2006), Henley et al. (2011), Vance et al. (2015) and Henley et al. (2017) may not identify meaningful decadal phases. ~~the above mentioned methods may not identify meaningful decadal phases.~~ Extraordinary long run lengths are identified if the overall mean is used as a threshold with some runs lasting several centuries (see Figure 2). It is not clear whether these centennial trends are due to low frequency climate variability or due to unknown local factors affecting the proxy. With such reconstructions, one can either exclude them from analysis or filter out the centennial trends. Exclusion forgoes any useful information from the reconstructions. On the other hand, filtering out the centennial trends offers the prospect of extracting possibly useful information about run lengths.

We adopt the latter approach proposing a dynamic threshold framework to filter out centennial trends. In this framework, potential crossings are taken to be the change points detected by the Mann-Whitney test, and the sign of PDV phases is determined by a dynamic threshold that takes centennial trends into consideration - this relaxes the restriction of a static threshold defining a phase crossing. Figure 2 provides insight about the mechanics of dynamic threshold method showing the PDV time series and the resulting block phase waveform. The key steps of this framework are:

1. Detect step-change points. For a given reconstruction, apply a change point detection method. A number of methods can be used. In this study we used the non-parametric Mann-Whitney test method (Mauget, 2003) with a given window width w and confidence level α to identify the step-change points. The method involves centring a window of width w at a particular year t and then applying the Mann-Whitney test to the samples of width w/2 at or before and after year t. A step change is deemed to occur if the Mann-Whitney test statistic lies outside the α confidence limits (under the null hypothesis). ~~For a given reconstruction, apply a change point detection method. This study will use the Mann-Whitney test method (Mauget, 2003) with a given window width *w* and confidence level α to identify the step-change points. However, other methods can be used.~~

2. Merge consecutive step-change points. When two or more step change points occur in consecutive years, replace them with a single change point. That is, if there are either 2*n-1* or 2*n* years that are determined as consecutive change points (*n*=1, 2, …), replace them with one change point at year *n*. This guarantees that the new change point is always closest to the middle of the run of consecutive step-change points.

3. Assign a phase to each run defined by the interval bounded by two step-change points. Let $i(t)$ denote the PDV index in year t, $i_f(t)$ the filtered PDV index using a first-order Butterworth filter (which was used by Henley et al. (2011) to filter paleo PDV indices)~~(Selesnick and Burrus, 1998)~~ with cut-off frequency 1/*y* year^{-1}, t_{cp}^k the year that the k^{th} change point occurs (from steps 1 and 2) and *s(t)* the phase state of year t. The mean PDV index for the k^{th} run is

$$\bar{i}^k = \frac{1}{(t_{cp}^k - t_{cp}^{k-1} + 1)} \sum_{t_{cp}^{k-1}}^{t_{cp}^k} i(t) \quad (1)$$

while the corresponding mean of the filtered index is

$$\bar{i}_f^k = \frac{1}{(t_{cp}^k - t_{cp}^{k-1} + 1)} \sum_{t_{cp}^{k-1}}^{t_{cp}^k} i_f(t) \quad (2)$$

The phase of the k^{th} run length is then defined by

$$s(t) = \begin{cases} 1 \text{ if } \bar{i}^k \geq \bar{i}_f^k \\ 0 \text{ if } \bar{i}^k < \bar{i}_f^k \end{cases}, t \in (t_{cp}^{k-1}, t_{cp}^k - 1) \quad (3)$$

4. Determine run lengths using the time series *s(t)*. The run length of the k^{th} run will be $t_{cp}^k - t_{cp}^{k-1}$.

This dynamic threshold method can be seen as a generalization of previous run length extraction methods. If parameter *y* is set to the total number of years in a given reconstruction, the

dynamic threshold method reduces to the method used in Verdon and Franks (2006). If change points are defined by the PDV index crossing the threshold and the cut-off frequency parameter *y* of the first-order Butterworth filter is set to the total number of years in the reconstruction, the dynamic threshold reduces to the method used in Henley et al. (2011) and Henley et al. (2017).

Three parameters need to be specified in the dynamic threshold method: window width *w* and confidence level α in the Mann-Whitney test, and cut-off frequency *y* of the Butterworth filter. ~~To filter out centennial trends, *y* should be set large enough to account for the centennial trend.~~The standard cut-off frequency parameter in the dynamic threshold method is used to filter out centennial trends that may be mixed with decadal variability and should have little influence on the statistical characteristics of the inferred runs. A cut-off frequency of 1/100 years is considered to adequately meet these requirements. Hence, *y=100* is used in this study. In this study, we set the confidence level to be 90%, a value that is consistent with other reconstruction studies (eg Gedalof and Smith, 2001; Shen et al., 2006; McGregor et al., 2010; Pent et al., 2015). It has been shown that reconstructions are sensitive to the choice of the window used in the Mann-Whitney test as well as the choice of threshold (Henley et al., 2017;Henley, 2013). Accordingly, this study investigates the sensitivity of dynamic threshold method results to the choice of window width *w* in the Mann-Whitney test. This sensitivity analysis is used to guide the selection of parameters to be used in the analysis of all 12 PDV reconstructions.

3.2 Results from different run length extraction methods

3.2.1 Comparison of static and dynamic threshold

To further demonstrate the shortcomings of the static threshold method and the need for a dynamic threshold method, run length samples are extracted using the dynamic threshold method proposed in this study (denoted as “Dynamic” in the figures) and the static threshold method used in

Verdon and Franks (2006) (denoted as "Static" in the figures). These two methods only differ in step 3 of Section 3.1. In the static method, the overall mean defined as $\overline{i_f} = \frac{1}{n}\sum_{1}^{n} i_f(t)$ (in which n is number of years in corresponding reconstruction) is used as the threshold instead of the dynamic threshold defined in equation (2). Extracted run length samples using the dynamic and static threshold methods from 3 millennium long reconstructions with centennial trends are plotted in Figure 2. Black lines are filtered standardized PDV indices; green lines represent the Butterworth (1/100) filtered data in the dynamic plots and the overall mean in the static plots; red lines represent the run extracted using either the dynamic or static thresholds method. The run length distributions are plotted in Figure 3.

It is shown in Figure 2 that extraordinarily long runs are identified when the static threshold method is used - the stronger the centennial trends, the longer the runs. On the other hand, use of the dynamic threshold framework appears to filter centennial trends and produce more meaningful run lengths. The run length density plots show that the dynamic threshold method for the Macd05 and Mann09 reconstructions removes centennial-scale runs. In the case of Vanc15 the centennial trends are muted, possibly because Vanc15 is annually resolved and very accurately dated (Vance et al., 2015) and as such is more likely to exhibit more realistic annual to decadal variability than Macd05 and Mann09 (see also Figure 2). Individual reconstructions are subject to different error structures and magnitudes. Use of multiple reconstructions can help reduce the impact of errors, and thus provide more insight into the statistical characteristics of PDV run lengths. Nonetheless, errors still need to be carefully considered. This is illustrated by Henley et al. (2011) who combines individual reconstructions according to their accuracy. However, the accuracies of the individual reconstructions are not high (Nash-Sutcliffe indices <0.48) suggesting that even favouring the more accurate individual reconstructions will still result in a combined/overall reconstruction with large errors. Therefore, the focus of our research is to identify a signal (or signals) that is (are) common to all individual reconstructions.

3.2.2 Window width sensitivity analysis

Henley (2013) demonstrated that the mean run length is strongly dependent on the choice of window width in Mann-Whitney method and concluded that a subjective choice of window width is likely to bias the resulting run length distributions. To explore the sensitivity of window width on run length distributions, run length boxplots for all reconstructions are plotted in Figure 4 for window widths from 10 to 60 years with step of 2, with positive and negative samples plotted together and separately.

Figure 4 shows that the run length distributions clearly depend on the choice of window width with longer window widths leading to longer run length samples. However, the change is gradual particularly for shorter window widths. As a result, run length distributions are not particularly sensitive to small variations in window width of the order of 10 years. ~~That is, runs extracted by using 20-year window width are different from using 50-year window width, but remain similar to using 30-year window width.~~

3.2.3 Determination of parameters and application of dynamic threshold method

To guide the selection of an operationally useful window width, several window widths are applied to the three instrumental PDV time series to identify which produces results that match existing knowledge about run length in instrumental periods. Four window widths, 10, 20, 30 and 40 years, are applied to the instrumental time series with the extracted runs plotted in Figure 5. Interpretation of Figure 5 is illustrated for Figure 5d. The green line represents the IPO_Henley index, from which three runs, denoted by green circles, are extracted: ~1900-1938 (positive), ~1938-1976 (negative), ~1976-2004 (positive). These are very similar (in number, sign and duration) to the black (PDO_Mantua) and red run lengths (IPO_Parker). It can be seen that a window width of 10 years identifies very short runs which are more likely to represent random fluctuations rather than different

PDV phases - we note that Henley et al. (2017) discarded runs with lengths less than 5 years. However, for the longer window widths of 20, 30 and 40 years, the differences become more muted.

To better understand these differences, consider window widths of 20 and 40 years. When the 40-year window is used, the first 20 years of the window are compared to the last 20 years. As a result, runs of less than 20 years may be overlooked. To demonstrate this, different window widths are applied to all reconstructions to extract run lengths, from which conditional run length distributions are determined. Figure 6 presents run length distributions in six panels. Panels (a1) to (a3) compare distributions for 20 and 40-year windows. Panel (a1) shows the empirical density of run lengths, while panels (a2) and (a3) show the empirical densities conditioned on run lengths less than 20 years and greater than 20 years respectively. Likewise Panels (b1) to (b3) repeat this sequence for 20 and 60-year windows with the conditioning threshold of 30 years. ~~Panel (a1) shows the unconditional run distribution, while panel (a2) shows the distribution conditioned on runs less than 20 years and panel (a3) shows the distribution conditioned on runs greater than 20. Panels (b1) to (b3) repeat this for 20 and 60-year windows with conditioning on the 30-year length.~~

Panels (a1) and (b1) show that the unconditional run length distributions are different using different window widths with the difference greater for a greater difference in window width. However, panels (a3) and (b3) show that the conditional distributions for the longer runs are very similar. This shows that both the short and long window widths extract the longer runs in a largely consistent manner. However, the choice of window width strongly affects the distribution of shorter runs as shown in panels (a2) and (b2). Overall, this demonstrates that shorter window widths can “see” higher frequency runs better than longer window widths but both “see” the same distribution of lower frequency runs.

These considerations suggest that the window width should be as small as possible, yet sufficiently big to filter out runs that are considered more likely to be random fluctuations rather than PDV phases. Therefore, in this study, a 20-year window is used to ensure decadal or longer runs are identified. Figure 7 plots the extracted runs for the 12 reconstructions using the dynamic threshold

method with 90% confidence limits ~~confidence level~~ and 20-year window in the Mann-Whitney test. The PDV indices plotted in Figure 7 were filtered using a Butterworth filter with a cut-off frequency of 1/11 years^{-1} to more clearly present the decadal variability in each reconstruction. In the case of Verd06, no PDV indices are available, so the original runs are plotted. Visual inspection of Figure 7 suggests that the identified runs in all reconstructions seem largely consistent with the multi-decadal fluctuations in the filtered PDV indices. In the case of reconstructions with pronounced centennial trends (such as Macd05 and Mann09), positive and negative runs tend to match multi-decadal fluctuations in the filtered indices even when the indices persist above or below the long-term average over centennial scales.

4. ~~Hydrologically important~~ Sstatistical signatures ~~associated with~~of PDV run lengths

The dynamic threshold method allows run lengths to be extracted from all the PDV reconstructions, even from those with centennial trends. If we had a perfect reconstruction there would be no need for any further reconstructions. However, the fact is that each paleoclimate reconstruction is subject to errors, both random and systematic, that are not fully understood. Therefore, it is pertinent to identify statistical features that are common to the reconstructions and also important to hydroclimatic risk assessments. This addresses the second objective of this study. In particular, the analysis seeks to answer the following questions: Are the distributions of positive and negative PDV run lengths statistically different, and is the variability seen in the instrumental record (the last ~100 years), or the shorter 300-400 year paleoclimate reconstructions, representative of the full range of variability that has occurred in the past or of what is plausible in the future? These questions are of particular importance for predicting near-term PDV phase behaviour and assessing near-term flood and drought risk and therefore are the primary focus of this section.

4.1 Are run lengths different during positive and negative PDV phases?

Vance et al. (2015) analysed an IPO reconstruction from AD 1000 to 2003 and concluded that the positive IPO phase (IPO>0.5) has an average duration of 14 years and the negative IPO phase (IPO< −0.5) has an average duration of 9 years. This is the first known analysis that considered positive and negative PDV phases separately.

To explore the hypothesis that run lengths have different distributions during positive and negative PDV phases, boxplots of run lengths for each reconstruction are shown in Figure 8. Panel (a) shows boxplots for all run lengths for each reconstruction, Panel (b) shows boxplots of pooled positive and negative run lengths, while Panel (c) shows boxplots of positive and negative run lengths for each reconstruction. Panel (b) suggests that the pooled median positive and negative run lengths are similar, but that positive run lengths are more variable. However, Panel (c) shows that there is considerable variability between reconstructions. The absence of a consistent difference between positive and negative run distributions may be a consequence of sampling variability masking any underlying difference. Sampling variability refers to the variability in the target statistics that arises when random sampling is repeated. Therefore, it is important to recognize that differences in run statistics may be due to sampling variability. It is only when differences are bigger than what would be expected due to sampling variability that we would conclude there is a significant difference. This seems reasonable given that the average number of phases (either positive or negative) per reconstruction is only 13. A more formal statistical analysis is required to explicitly deal with sampling variability arising from the small samples.

Henley et al. (2011) adopted the Gamma distribution as the probability model of PDV run lengths. The Gamma distribution has two parameters which are related to the mean μ and standard deviation σ of the run lengths. For each reconstruction, the positive and negative runs are extracted and the posterior distribution of the mean and standard deviation is inferred using MCMC (Gelman

and Rubin, 1992). Figure 9 presents the joint posterior distribution of the differences in mean ($\mu^+ - \mu^-$) and standard deviation ($\sigma^+ - \sigma^-$) for each reconstruction. If the difference between positive and negative runs is statistically significant, one would expect the posterior distribution to be well removed from the zero-difference origin.

Inspection of Figure 9 suggests there is no strong and consistent evidence to reject the assumption that positive and negative phase run length distributions are the same. Although the mean of positive phase run lengths tends to be longer than negative phase run lengths in several reconstructions (e.g. Vanc15, Henl11, Lins08, Darr01), the statistical significance of this difference is weak and the phenomenon is not consistent across all reconstructions. A similar conclusion applies to the standard deviation of positive and negative phase run lengths. This highlights the fundamental limitation of the small PDV run samples derived from the reconstructions.

4.2 Are run lengths stationary over the last millennium?

With paleoclimate reconstructions becoming longer, questions arise whether the PDV run length has been stationary in the past millennium and whether a shorter paleoclimate reconstruction is representative of the full range of variability that has occurred in the past or of what is plausible in the future. The stationarity issue has been explored most PDV reconstructions used in this study. Biondi et al. (2001) identified weakened amplitude of bi-decadal oscillations in the late 1700s to mid-1800s. Based on reconstruction over 1700-1979, D'Arrigo et al. (2001) declared that variations at around 12-17 years are considerably more pronounced from 1700-1849 relative to 1850-1979 with a shift towards decreased amplitude since about 1850. Using visual inspection, Gedalof and Smith (2001) stated that the 30-70 year PDV frequency is confined to the pre-1840 portion of the series over the period of 1599-1983. D'Arrigo and Wilson (2006) reported a broad range of lower (multi-decadal to centennial) frequencies over the 1565-1988 period. Shen et al. (2006) pointed out that the major PDO regime timescale modes of oscillation have not been persistent over 1470-1998 and that 75-115 year and 50 –70 year oscillations dominated the periods before and after 1850, respectively. Linsley et al.

(2008) argued that decadal to inter-decadal variability in the south Pacific convergence zone region has been relatively constant over 1650-2004. MacDonald and Case (2005) presented evidence for a strong and persistent negative PDO state during the medieval period (AD 900 to 1300), suggesting a cool north eastern Pacific at that time. Mann et al. (2009) observed that the Little Ice Age (LIA, 1400 to 1700) and the Medieval Climate Anomaly (MCA, 950 to 1250) showed extra variability and persistence, challenging the assumption of a stationary climate in the past millennium. The mechanism responsible for the extra variability and persistence in studies done by MacDonald and Case (2005) and Mann et al. (2009) is unclear and may be explained by centennial climate variability. It is unknown whether the PDV structure in the past millennium has remained stationary during periods that display extra variability and persistence. This issue can be investigated by filtering out the centennial trends.

All the above stationarity studies are based on single reconstructions. A further complication arises for reconstructions that only cover the last ~300-400 years where the underlying non-stationarity may be masked as a consequence of sampling variability or the sampling variability is misinterpreted as non-stationarity. To mitigate this, here we use only millennium-length records, Macd05 (993-1996), Mann09 (500-2006) and Vanc15 (1000-2003), in an attempt to obtain statistically meaningful conclusions on PDV stationarity. The millennium-length records are split into pre-1600 and post-1600 samples to explore whether run length characteristics are statistically similar in these two periods and to shed light on whether shorter records (either the ~100 year instrumental record or the 300-400 year paleoclimate reconstructions) are able to represent the full range of variability. The year 1600 is selected as the change point because most of the shorter PDV reconstructions date back to ~1600 (refer to Table 1 and Figure 2 for details).

Figure 10 presents boxplots of pre-1600 and post-1600 run lengths: panel (a) presents boxplots of pooled pre-1600 and post-1600 run lengths, while panel (b) shows boxplots of pre-1600 and post-1600 runs for each of the three millennium-length reconstructions. A consistent pattern

emerges in which the median run length and interquartile range are greater in the pre-1600 period for both pooled samples (Figure 10a) and each individual reconstruction (Figure 10b).

Bearing in mind that the samples are small, a Gamma distribution was inferred for the pre- and post-1600 samples. Figure 11 presents the joint posterior distribution of the differences in the pre- and post-1600 means and standard deviations for each reconstruction using all run lengths. A consistent pattern once again emerges. For all three reconstructions the posteriors of the pre- and post-1600 differences lie in the lower left quadrant with the origin minimally intersecting with the posterior. Therefore, the evidence that there is a difference is visually strong.

4.3 Discussion

~~All of the reconstructions in Table 1 have focused on the run length of PDV, with positive and negative PDV run lengths being pooled together in their analysis, except for Vance et al. (2015).~~ All of the reconstructions reported in Table 1 did not distinguish between positive and negative PDV runs in their analysis, except for Vance et al. (2015). Based on a millennium IPO reconstruction, Vance et al. (2015) found that IPO has an average positive phase (IPO>0.5) duration of 14 years and a negative phase (IPO< −0.5) duration of 9 years over A.D.1000-2003, and concluded that positive and negative phases durations and frequencies were different in the last millennium. This is the first study that addresses positive and negative PDV runs separately. However, based on comparisons with other reconstructions our results show that there remains uncertainty as to whether or not the run lengths of positive epochs are statistically different to the run length of negative epochs. It is important to keep in mind that the analysis here is fundamentally limited by small PDV run length sample sizes from most of the reconstructions. Two theories may explain the absence of a consistent difference between positive and negative run distributions. One is that no statistically meaningful differences exist and the other is that differences do exist but are masked by sampling variability. The latter seems plausible given the average number of phases (either positive or negative) per reconstruction are only 13.

All three millennium-length records lead to higher inferred run length mean and standard deviation in the pre-1600 samples, suggesting longer and more varied PDV runs during the period before 1600 AD. The fact that the differences in the mean and standard deviation of run lengths pre- and post-1600 appear to be statistically significant raises an important question, should the information from pre-1600 reconstructions be used to infer PDV behaviour in the near climate. If one adopts a gradualist view of climate non-stationarity, the immediate past would be considered more representative of the present than the more distant past. Setting aside for the moment the possibility that the climate-proxy relationship may be not be stationary, the gradualist perspective would support discarding the information from the pre-1600 reconstruction on the grounds that it introduces bias when inferring a statistical model of near climate PDV runs. We offer two reasons opposing this perspective.

First, there is no assurance that climate non-stationarity evolves in a gradual manner so that PDV behaviour in the near climate is better represented by the post-1600 climate than the pre-1600 climate. Indeed there is evidence to the contrary, namely that non-stationarity may be characterized by seemingly shorter-term shifts. Ho et al. (2017) found that American streamflow in the twentieth century featured longer wet and dry spells compared to the preceding 450 years, suggesting the possibility of occurrence of extended dry or wet periods in the future that exceeds variability presented in instrumental and short (i.e. < 500 years) paleoclimate reconstructions. Similar conclusions are also drawn in other studies (Vance et al., 2015;Tozer et al., 2016;Ho et al., 2015b, a).

Second, the inclusion of pre-1600 records offers a significantly larger sample of run lengths and therefore is more likely to capture the full range of what is plausible. Extended wet/dry periods considered as rare and extreme (or even impossible) based on recent (i.e. instrumental) history might actually be more likely than thought (or even common) when put into context of climate conditions seen over the last 1000-2000 years. Even if one accepts the gradualist view of non-stationarity, the increased information about PDV more than outweighs the potential bias arising from the use of an apparently statistically inconsistent record. At least until such time as it is proven that the climate has

totally shifted and that what occurred in the past is no longer possible – also required is identification of when that shift occurred.

It is therefore clear that in addition to short instrumental records inadequately sampling the length and severity of dry/wet epochs, the shorter paleoclimate reconstructions may similarly misrepresent hydroclimatic variability and persistence. For instance, from Ho et al. (2015a), when using the period 1684–1980, both the driest and wettest instrumental decadal rainfall lies closer to the middle of the minimum and maximum decadal reconstructed rainfall range. However, using the same method, Ho et al. (2015a) found that when 2751 years of reconstructions were examined, the wettest instrumental decadal rainfall is near the bottom of the maximum decadal reconstructed rainfall range, while the driest instrumental decadal rainfall is near the top of the minimum decadal reconstructed rainfall range. This suggests that statistics inferred from either instrumental data or short paleoclimate reconstructions will underestimate the variability contained in the longer palaeoclimate reconstructions. Therefore, research focused on developing and analysing information about hydroclimatic conditions over the last 1000 years or more (e.g. multi-millennium paleoclimate reconstructions) is critical if we are to better understand, quantify and manage the full range of hydroclimatic variability, and associated risks, that are possible in the future.

This study is affected by two fundamental limitations. One is the small PDV run sample size from the reconstructions. For instance, the average number of phases (either positive or negative) per reconstruction are only 13. Another limitation is that all statistical signatures are based on paleoclimate reconstructions. Although multiple reconstructions are used, this study is constrained by the intrinsic limitations of paleoclimate reconstruction based interpretations, such as assumptions of stationarity of the proxy-climate relationship (Phipps et al., 2013). The possibility that all these reconstructions are biased in the same direction cannot be ruled out and as such conclusions based on multiple reconstructions may also be biased.

5. Conclusions

PDV, and associated run lengths of predominantly dry or wet conditions, has profound implications for precipitation/streamflow prediction, flood/drought risk assessment and water resource management. Therefore, a better understanding of the statistical characteristics of PDV is needed. However, because instrumental records are short (~100 years at best in Australia), there is considerable uncertainty about the key statistical signatures of PDV, including run lengths. Paleoclimate reconstructions, serving as a vehicle to provide longer realizations of PDV, have been widely studied and used. However, for various reasons (e.g. proxy sources, site locations, proxy resolutions, reconstruction methods and local non-PDV effects) temporal coherence between different reconstructions varies significantly (Kipfmueller et al., 2012). Hence, one PDV reconstruction may lead to significantly different conclusions from another reconstruction.

The aim of this paper was to explore the characteristics of key statistical signatures of PDV based on multiple PDV reconstructions. We focused on the duration of dry or wet hydrological epochs (i.e. run lengths) as the key statistical signature and developed a robust method for extracting run lengths from multiple reconstructions. Extracting run lengths objectively is challenging, given interactions with sources of variability on a variety of temporal and spatial scales. For instance, when millennium-length reconstructions are used, variability and persistence at the centennial scale can lead to biased characterisation of decadal climate variability. The dynamic threshold framework introduced in this study, which takes centennial trends into consideration, was shown to extract meaningful run length information from multiple reconstructions.

No strong evidence was found to support the assumption that run lengths have statistically different distributions during positive and negative PDV phases. Analysis based on three millennium long reconstructions suggests that it is more likely than not that PDV run length has been non-stationarity in the past millennium. This again highlights that the instrumental record (~100 years at best), and even short paleoclimate reconstructions (i.e. less than 400 years into the past), should not be

assumed to represent the full range of variability that has occurred in the past or what may occur in the future. Caution should be exercised regarding assumptions that the climate is stationary and the implications of a non-stationary climate should at least be tested. Longer climate reconstructions (i.e. 1000 years or more) appear to give more useful information and insights into what has occurred in the past and also tell us more about what is plausible and what needs to be planned for in the future.

All of the reconstructions explored in this study have focused on the run length of PDV, yet they have provided quite different run length characterizations. This again highlights that each reconstruction is subject to errors, both random and systematic, that are not fully understood. Therefore, a challenging and important research direction is the analysis of run length characteristics with explicit consideration of the errors. Another important research direction is the continued refinement of methods that utilize PDV information in water resource management as we work toward improving how we assess and manage hydrological risks in a variable and changing climate.

Acknowledgments

Funding for this research was provided by Australian Research Council Linkage Grant LP120200494 with further funding and/or in-kind support also provided by the NSW Office of Environment and Heritage, Sydney Catchment Authority, Hunter Water Corporation, NSW Office of Water, and NSW Department of Finance and Services.

References

Andreoli, R. V., and Kayano, M. T.: ENSO-related rainfall anomalies in South America and associated circulation features during warm and cold Pacific decadal oscillation regimes, International Journal of Climatology, 25, 2017-2030, 10.1002/joc.1222, 2005.

Atwood, A. R., Wu, E., Frierson, D. M. W., Battisti, D. S., and Sachs, J. P.: Quantifying Climate Forcings and Feedbacks over the Last Millennium in the CMIP5–PMIP3 Models, Journal of Climate, 29, 1161-1178, 10.1175/JCLI-D-15-0063.1, 2015.

Biondi, F., Gershunov, A., and Cayan, D. R.: North Pacific decadal climate variability since 1661, Journal of Climate, 14, 5-10, 10.1175/1520-0442(2001)014<0005:NPDCVS>2.0.CO;2, 2001.

Chiew, F. H. S.: Estimation of rainfall elasticity of streamflow in Australia, Hydrological Sciences Journal, 51, 613-625, 10.1623/hysj.51.4.613, 2006.
Cook, B. I., Smerdon, J. E., Seager, R., and Cook, E. R.: Pan-Continental Droughts in North America over the Last Millennium, Journal of Climate, 27, 383-397, 10.1175/JCLI-D-13-00100.1, 2013.
D'Arrigo, R., Villalba, R., and Wiles, G.: Tree-ring estimates of Pacific decadal climate variability, Clim Dyn, 18, 219-224, 10.1007/s003820100177, 2001.
D'Arrigo, R., and Wilson, R.: On the Asian expression of the PDO, International Journal of Climatology, 26, 1607-1617, 10.1002/joc.1326, 2006.
Dai, A.: The influence of the inter-decadal Pacific oscillation on US precipitation during 1923–2010, Clim Dyn, 41, 633-646, 10.1007/s00382-012-1446-5, 2013.
Dong, B., and Dai, A.: The influence of the Interdecadal Pacific Oscillation on Temperature and Precipitation over the Globe, Clim Dyn, 45, 2667-2681, 10.1007/s00382-015-2500-x, 2015.
Folland, C. K., Renwick, J. A., Salinger, M. J., and Mullan, A. B.: Relative influences of the interdecadal Pacific oscillation and ENSO on the South Pacific convergence zone, Geophysical Research Letters, 29, 1643, 10.1029/2001GL014201, 2002.
Gedalof, Z., and Smith, D. J.: Interdecadal climate variability and regime-scale shifts in Pacific North America, Geophysical Research Letters, 28, 1515-1518, 10.1029/2000GL011779, 2001.
Gelman, A., and Rubin, D. B.: Inference from Iterative Simulation Using Multiple Sequences, Statistical Science, 7, 457-472, 10.1214/ss/1177011136, 1992.
Goodrich, G. B., and Walker, J. M.: The Influence of the PDO on Winter Precipitation During High- and Low-Index ENSO Conditions in the Eastern United States, Physical Geography, 32, 295-312, 10.2747/0272-3646.32.4.295, 2011.
Henley, B., Gergis, J., Karoly, D., Power, S., Kennedy, J., and Folland, C.: A Tripole Index for the Interdecadal Pacific Oscillation, Clim Dyn, 1-14, 10.1007/s00382-015-2525-1, 2015.
Henley, B. J., Thyer, M. A., Kuczera, G., and Franks, S. W.: Climate-informed stochastic hydrological modeling: Incorporating decadal-scale variability using paleo data, Water Resources Research, 47, W11509, 10.1029/2010WR010034, 2011.
Henley, B. J., Thyer, M. A., and Kuczera, G.: Climate driver informed short-term drought risk evaluation, Water Resources Research, 49, 2317-2326, 10.1002/wrcr.20222, 2013.
Henley, B. J.: Pacific decadal climate variability: Indices, patterns and tropical-extratropical interactions, Global and Planetary Change, 155, 42-55, 10.1016/j.gloplacha.2017.06.004, 2017.
Henley, B. J., Meehl, G., Power, S., Folland, C., King, A., Brown, J., Karoly, D., Delage, F., Gallant, A., and Freund, M.: Spatial and temporal agreement in climate model simulations of the interdecadal pacific oscillation, Environmental Research Letters, 12, 044011, 10.1088/1748-9326/aa5cc8, 2017.
Ho, M., Kiem, A. S., and Verdon-Kidd, D. C.: A paleoclimate rainfall reconstruction in the Murray-Darling Basin (MDB), Australia: 2. Assessing hydroclimatic risk using paleoclimate records of wet and dry epochs, Water Resources Research, 51, 8380-8396, 10.1002/2015WR017059, 2015a.
Ho, M., Kiem, A. S., and Verdon-Kidd, D. C.: A paleoclimate rainfall reconstruction in the Murray-Darling Basin (MDB), Australia: 1. Evaluation of different paleoclimate archives, rainfall networks, and reconstruction techniques, Water Resources Research, 51, 8362-8379, 10.1002/2015WR017058, 2015b.
Ho, M., Lall, U., Sun, X., and Cook, E. R.: Multiscale temporal variability and regional patterns in 555 years of conterminous U.S. streamflow, Water Resources Research, 3047–3066, 10.1002/2016WR019632, 2017.
Hoell, A., and Funk, C.: Indo-Pacific sea surface temperature influences on failed consecutive rainy seasons over eastern Africa, Clim Dyn, 43, 1645-1660, 10.1007/s00382-013-1991-6, 2014.
Hu, Z.-Z., and Huang, B.: Interferential Impact of ENSO and PDO on Dry and Wet Conditions in the U.S. Great Plains, Journal of Climate, 22, 6047-6065, 10.1175/2009JCLI2798.1, 2009.
Kiem, A. S., Franks, S. W., and Kuczera, G.: Multi-decadal variability of flood risk, Geophysical Research Letters, 30, 1035, 10.1029/2002GL015992, 2003.

Kiem, A. S., and Franks, S. W.: Multi-decadal variability of drought risk, eastern Australia, Hydrological Processes, 18, 2039-2050, 10.1002/hyp.1460, 2004.
Kipfmueller, K. F., Larson, E. R., and St. George, S.: Does proxy uncertainty affect the relations inferred between the Pacific Decadal Oscillation and wildfire activity in the western United States?, Geophysical Research Letters, 39, PA2219, 10.1029/2011GL050645, 2012.
Kirtman, B., Power, S. B., Adedoyin, A. J., Boer, G. J., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A. M., Kimoto, M., and Meehl, G.: Near-term climate change: projections and predictability, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge,United Kingdom and New York, NY, USA, 953–1028, 2013.
Krishnan, R., and Sugi, M.: Pacific decadal oscillation and variability of the Indian summer monsoon rainfall, Clim Dyn, 21, 233-242, 10.1007/s00382-003-0330-8, 2003.
Li, L., Li, W., and Kushnir, Y.: Variation of the North Atlantic subtropical high western ridge and its implication to Southeastern US summer precipitation, Clim Dyn, 39, 1401-1412, 10.1007/s00382-011-1214-y, 2012.
Linsley, B. K., Zhang, P., Kaplan, A., Howe, S. S., and Wellington, G. M.: Interdecadal-decadal climate variability from multicoral oxygen isotope records in the South Pacific Convergence Zone region since 1650 AD, Paleoceanography, 23, PA2219, 10.1029/2007PA001539, 2008.
Ma, Z.: The interdecadal trend and shift of dry/wet over the central part of North China and their relationship to the Pacific Decadal Oscillation (PDO), Chinese Science Bulletin, 52, 2130-2139, 10.1007/s11434-007-0284-z, 2007.
MacDonald, G. M., and Case, R. A.: Variations in the Pacific Decadal Oscillation over the past millennium, Geophysical Research Letters, 32, L08703, 10.1029/2005GL022478, 2005.
Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ammann, C., Faluvegi, G., and Ni, F.: Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly, Science, 326, 1256-1260, 10.1126/science.1177303, 2009.
Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C.: A Pacific interdecadal climate oscillation with impacts on salmon production, Bulletin of the american Meteorological Society, 78, 1069-1079, 10.1175/1520-0477(1997)078<1069:APICOW>2.0.CO;2, 1997.
Mantua, N. J., and Hare, S. R.: The Pacific decadal oscillation, Journal of Oceanography, 58, 35-44, 10.1023/a:1015820616384, 2002.
Mauget, S. A.: Multidecadal regime shifts in US streamflow, precipitation, and temperature at the end of the twentieth century, Journal of Climate, 16, 3905-3916, 10.1175/1520-0442(2003)016<3905:MRSIUS>2.0.CO;2, 2003.
McCabe, G. J., Ault, T. R., Cook, B. I., Betancourt, J. L., and Schwartz, M. D.: Influences of the El Niño Southern Oscillation and the Pacific Decadal Oscillation on the timing of the North American spring, International Journal of Climatology, 32, 2301-2310, 10.1002/joc.3400, 2012.
McGregor, S., Timmermann, A., and Timm, O.: A unified proxy for ENSO and PDO variability since 1650, Climate of the Past, 6, 1-17, 10.5194/cp-6-1-2010, 2010.
Mehta, V. M., Rosenberg, N. J., and Mendoza, K.: Simulated Impacts of Three Decadal Climate Variability Phenomena on Water Yields in the Missouri River Basin1, JAWRA Journal of the American Water Resources Association, 47, 126-135, 10.1111/j.1752-1688.2010.00496.x, 2011.
Micevski, T., Franks, S. W., and Kuczera, G.: Multidecadal variability in coastal eastern Australian flood data, Journal of Hydrology, 327, 219-225, https://doi.org/10.1016/j.jhydrol.2005.11.017, 2006.
Parker, D., Folland, C., Scaife, A., Knight, J., Colman, A., Baines, P., and Dong, B.: Decadal to multidecadal variability and the climate change background, Journal of Geophysical Research: Atmospheres, 112, D18115, 10.1029/2007JD008411, 2007.
Phipps, S. J., McGregor, H. V., Gergis, J., Gallant, A. J. E., Neukom, R., Stevenson, S., Ackerley, D., Brown, J. R., Fischer, M. J., and Van Ommen, T. D.: Paleoclimate data–model comparison and the role of climate forcings over the past 1500 years, Journal of Climate, 26, 6915-6936, 10.1175/JCLI-D-12-00108.1, 2013.

Power, S., Casey, T., Folland, C., Colman, A., and Mehta, V.: Inter-decadal modulation of the impact of ENSO on Australia, Clim Dyn, 15, 319-324, 10.1007/s003820050284, 1999.
Razavi, S., Elshorbagy, A., Wheater, H., and Sauchyn, D.: Toward understanding nonstationarity in climate and hydrology through tree ring proxy records, Water Resources Research, 51, 1813-1830, 10.1002/2014WR015696, 2015.
Reason, C. J. C., and Rouault, M.: ENSO-like decadal variability and South African rainfall, Geophysical Research Letters, 29, 16-11-16-14, 10.1029/2002GL014663, 2002.
Saft, M., Western, A. W., Zhang, L., Peel, M. C., and Potter, N. J.: The influence of multiyear drought on the annual rainfall-runoff relationship: An Australian perspective, Water Resources Research, 51, 2444-2463, 10.1002/2014WR015348, 2015.
Selesnick, I. W., and Burrus, C. S.: Generalized digital Butterworth filter design, Trans. Sig. Proc., 46, 1688-1694, 10.1109/78.678493, 1998.
Shen, C., Wang, W. C., Gong, W., and Hao, Z.: A Pacific Decadal Oscillation record since 1470 AD reconstructed from proxy data of summer rainfall over eastern China, Geophysical Research Letters, 33, L03702, 10.1029/2005GL024804, 2006.
Tozer, C. R., Vance, T. R., Roberts, J. L., Kiem, A. S., Curran, M. A. J., and Moy, A. D.: An ice core derived 1013-year catchment-scale annual rainfall reconstruction in subtropical eastern Australia, Hydrol. Earth Syst. Sci., 20, 1703-1717, 10.5194/hess-20-1703-2016, 2016.
Vance, T. R., Roberts, J. L., Plummer, C. T., Kiem, A. S., and van Ommen, T. D.: Interdecadal Pacific variability and eastern Australian megadroughts over the last millennium, Geophysical Research Letters, 42, 129-137, 10.1002/2014GL062447, 2015.
Verdon, D. C., Wyatt, A. M., Kiem, A. S., and Franks, S. W.: Multidecadal variability of rainfall and streamflow: Eastern Australia, Water Resources Research, 40, W10201, 10.1029/2004WR003234, 2004.
Verdon, D. C., and Franks, S. W.: Long-term behaviour of ENSO: Interactions with the PDO over the past 400 years inferred from paleoclimate records, Geophysical Research Letters, 33, L06712, 10.1029/2005GL025052, 2006.
von Storch, H., Zorita, E., Jones, J. M., Gonzalez-Rouco, F., and Tett, S. F. B.: Response to Comment on ‘‘Reconstructing Past Climate from Noisy Data", Science, 312, 529, 10.1126/science.1121571, 2006.
Wahl, E. R., Ritson, D. M., and Ammann, C. M.: Comment on "Reconstructing Past Climate from Noisy Data", Science, 312, 529, 10.1126/science.1120866, 2006.
Wang, S., Huang, J., He, Y., and Guan, Y.: Combined effects of the Pacific Decadal Oscillation and El Niño-Southern Oscillation on Global Land Dry–Wet Changes, Scientific Reports, 4, 6651, 10.1038/srep06651
https://www.nature.com/articles/srep06651#supplementary-information, 2014.
Wu, Z., Huang, N. E., Long, S. R., and Peng, C.-K.: On the trend, detrending, and variability of nonlinear and nonstationary time series, Proceedings of the National Academy of Sciences, 104, 14889-14894, 10.1073/pnas.0701020104, 2007.
Zhang, Y., Wallace, J. M., and Battisti, D. S.: ENSO-like interdecadal variability: 1900-93, Journal of climate, 10, 1004-1020, 10.1175/1520-0442(1997)010<1004:ELIV>2.0.CO;2, 1997.

Tables

Table 1 Available PDV paleoclimate reconstructions and their descriptions

Reconstruction	Abbreviation	Period	Location	Proxy
Biondi et al. (2001)	Bion01	1661-1991	North Pacific (North America)	Tree ring based PDO
D'Arrigo et al. (2001)	Darr01	1700-1979	Northeast Pacific (Western North America)	Tree ring based PDO
Gedalof and Smith (2001)	Geda01	1599-1983	North Pacific (Coastal western North America)	Tree ring based PDO
MacDonald and Case (2005)	Macd05	993-1996	North Pacific (South California and Canada)	Tree ring based PDO
D'Arrigo and Wilson (2006)	Darr06	1565-1988	North Pacific (Eastern Asian)	Tree ring based PDO
Shen et al. (2006)	Shen06	1470-1998	Western North Pacific (Eastern China)	Proxy data of Summer Rainfall
Verdon and Franks (2006)	Verd06	1662-1998	Both North and South Pacific	Other reconstructions based PDO
Linsley et al. (2008)	Lins08	1650-2004	South Pacific (Fiji and Tonga)	Oxygen isotope from coral cores based IPO
Mann et al. (2009)	Mann09	500-2006	Global	Global tree ring, ice core, coral, sediment, and other assorted proxy records
McGregor et al. (2010)	Magr10	1650-1977	Global	low frequency variability of the proxies of ENSO
Henley et al. (2011)	Henl11	1471-2000 (filtered)	Both North and South Pacific	Other reconstructions
(Vance et al. (2015))	Vanc15	1000-2003 (filtered)	East Antarctica	Law Dome ice core based IPO

A list of figure captions

Figures

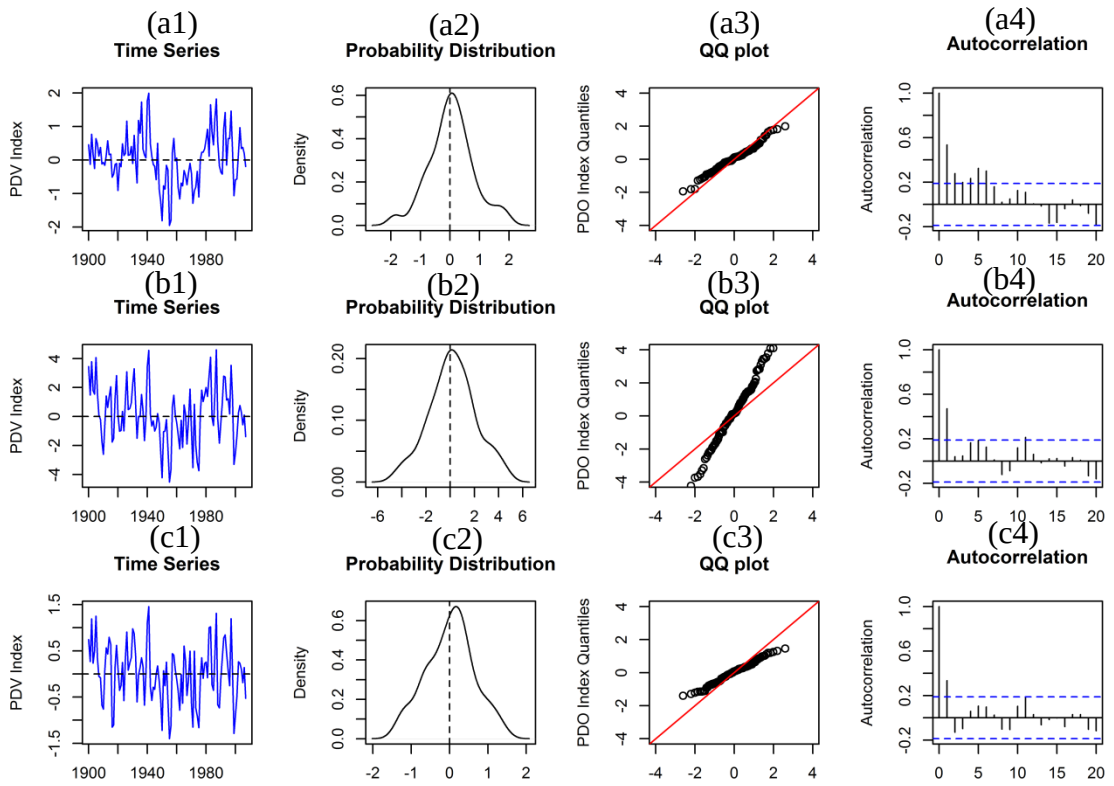


Figure 1 Annual instrumental time series, probability density plot, QQ plot and autocorrelation plot: (a): PDO from Mantua (1900-2015); (b): IPO from Parker (1871-2007); (c): IPO from Henley (1870-2007); The dashed lines in (a4)-(c4) present the 95% confidence bands for zero autocorrelation.

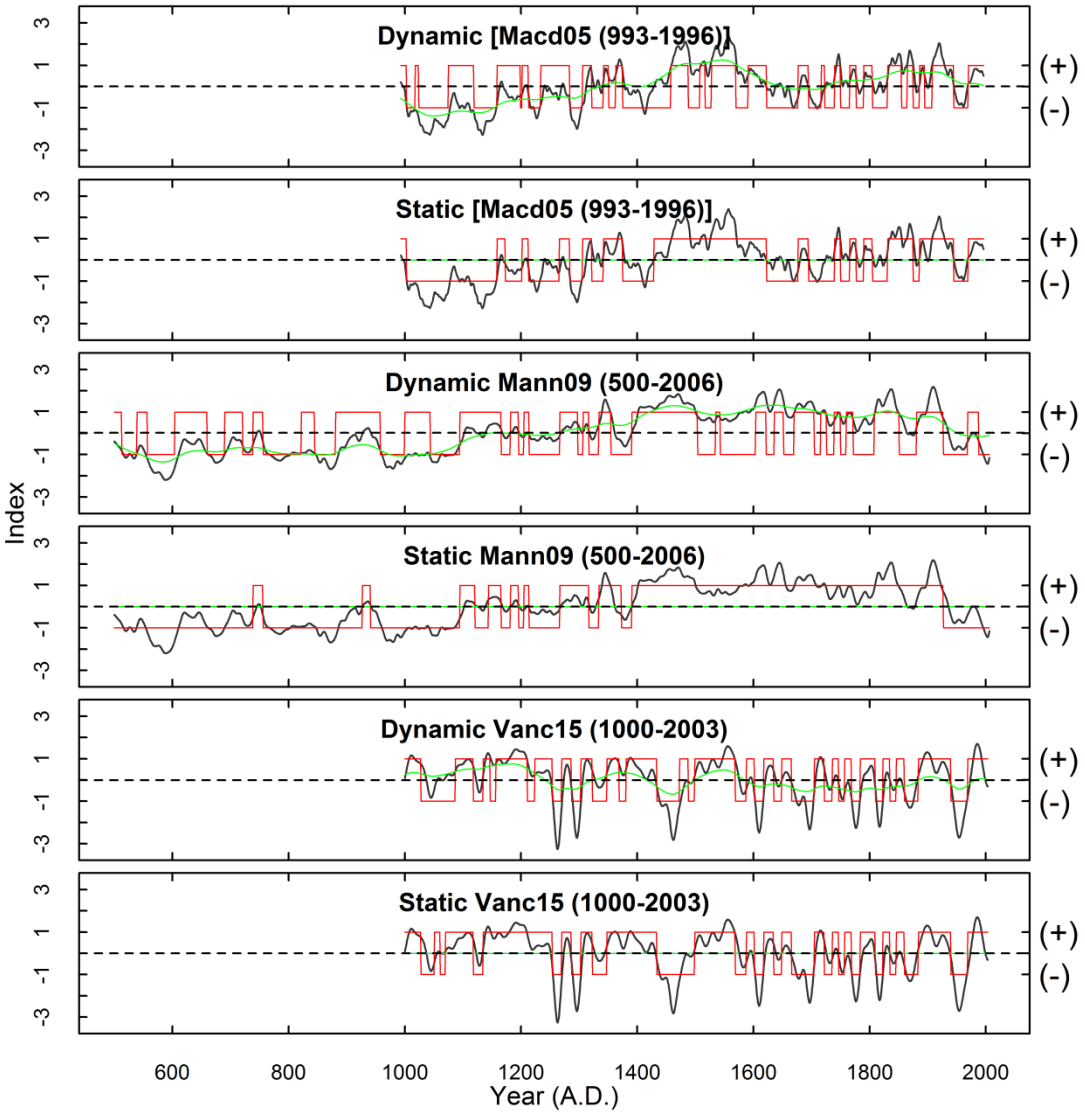


Figure 2 Run length time series extracted using the dynamic and static threshold methods for reconstructions with centennial trends

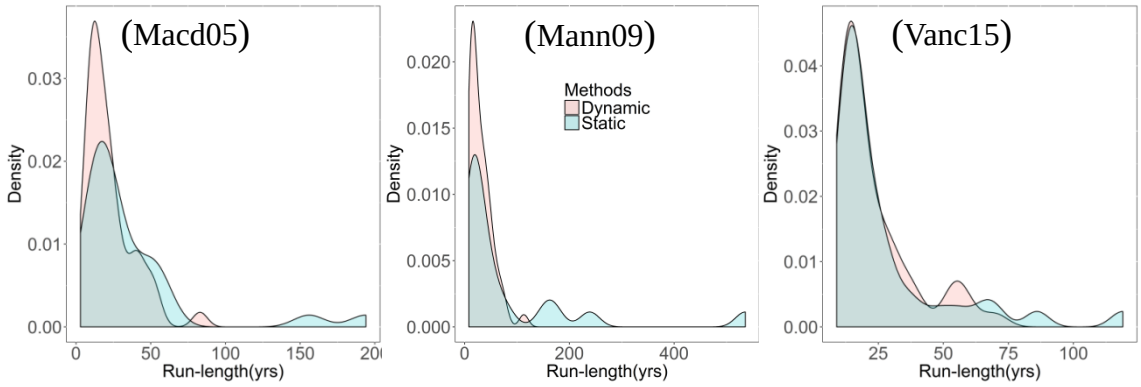


Figure 3 Comparison of run length distributions extracted using the dynamic and static threshold methods in reconstructions with centennial trends

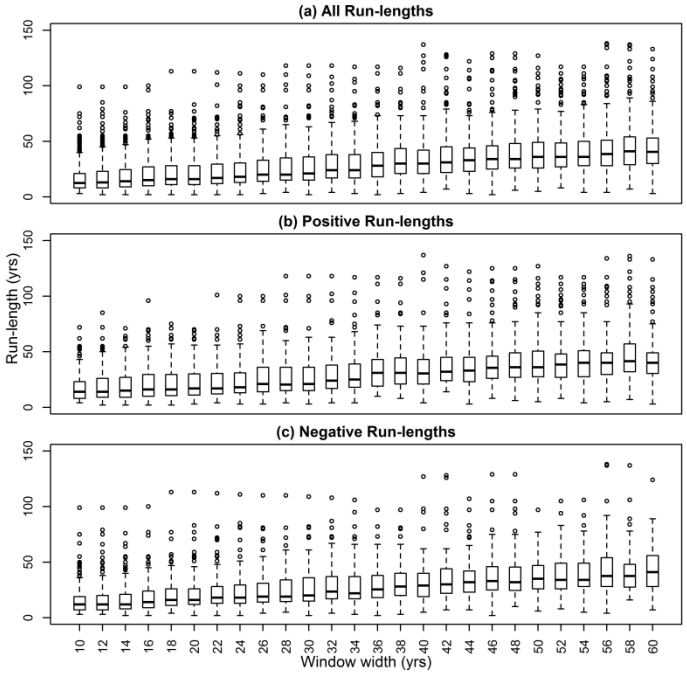


Figure 4 Boxplot of run lengths samples from all reconstructions with different window width in Mann-Whitney method

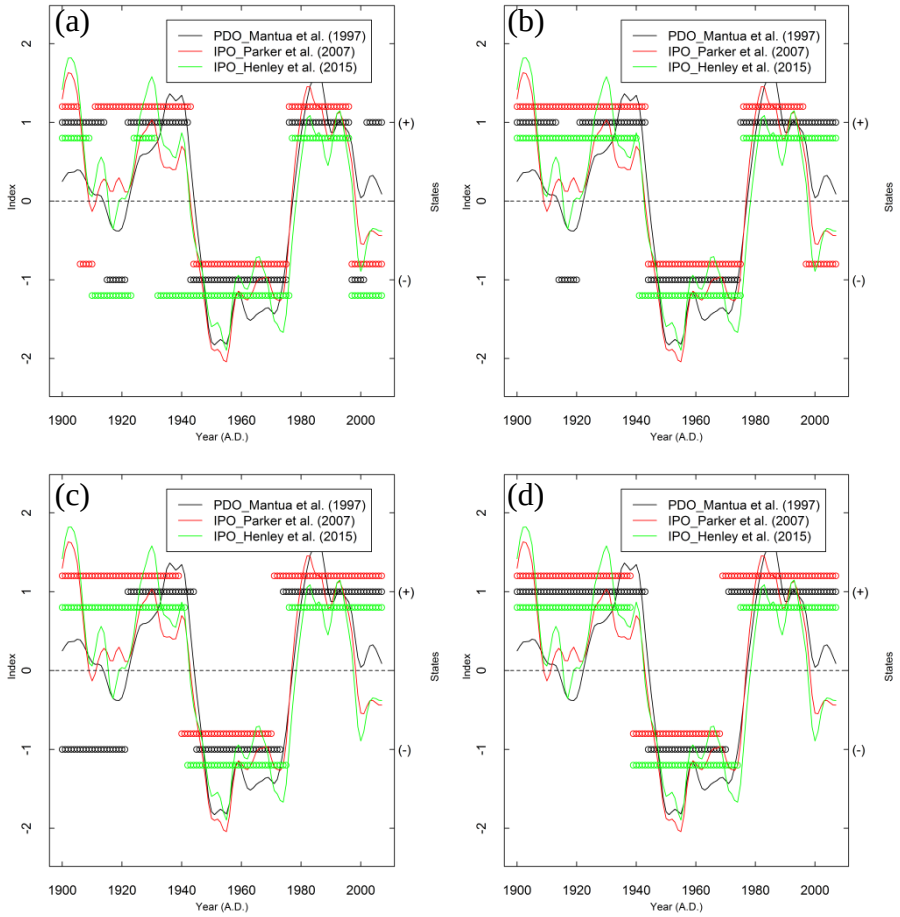


Figure 5 Time series plots of different instrumental PDV indices and corresponding run lengths (represented by dots in corresponding colour) extracted using different window: (a) 10; (b) 20; (c) 30; (d) 40 years

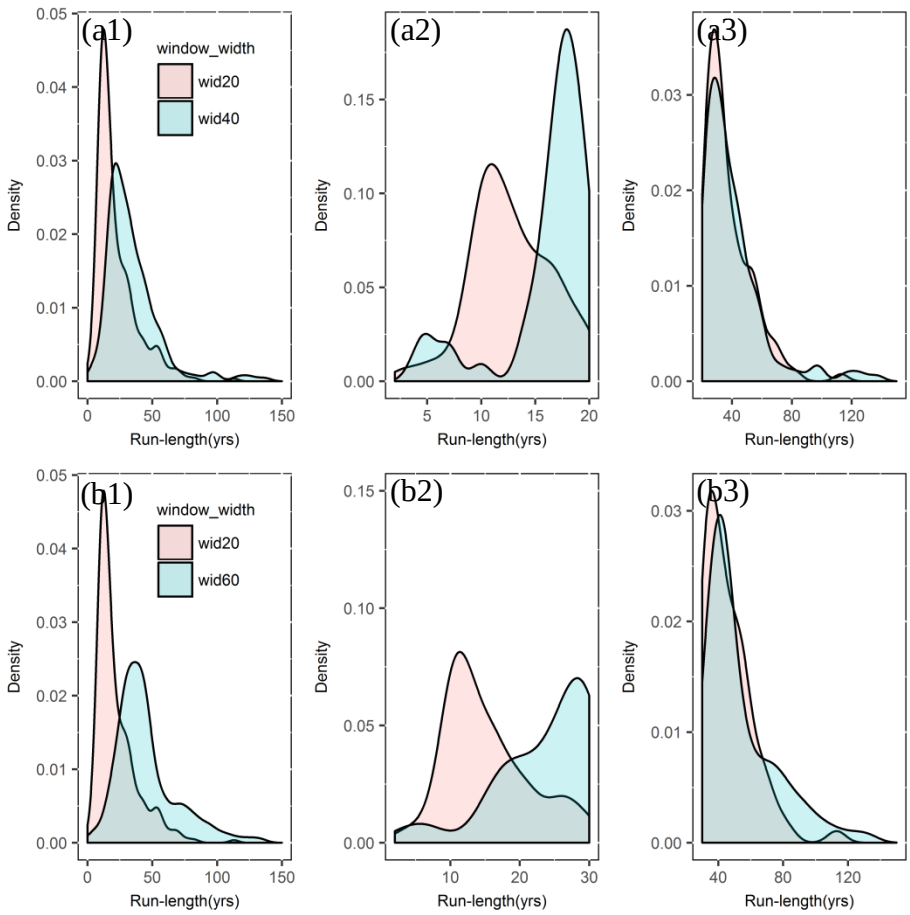
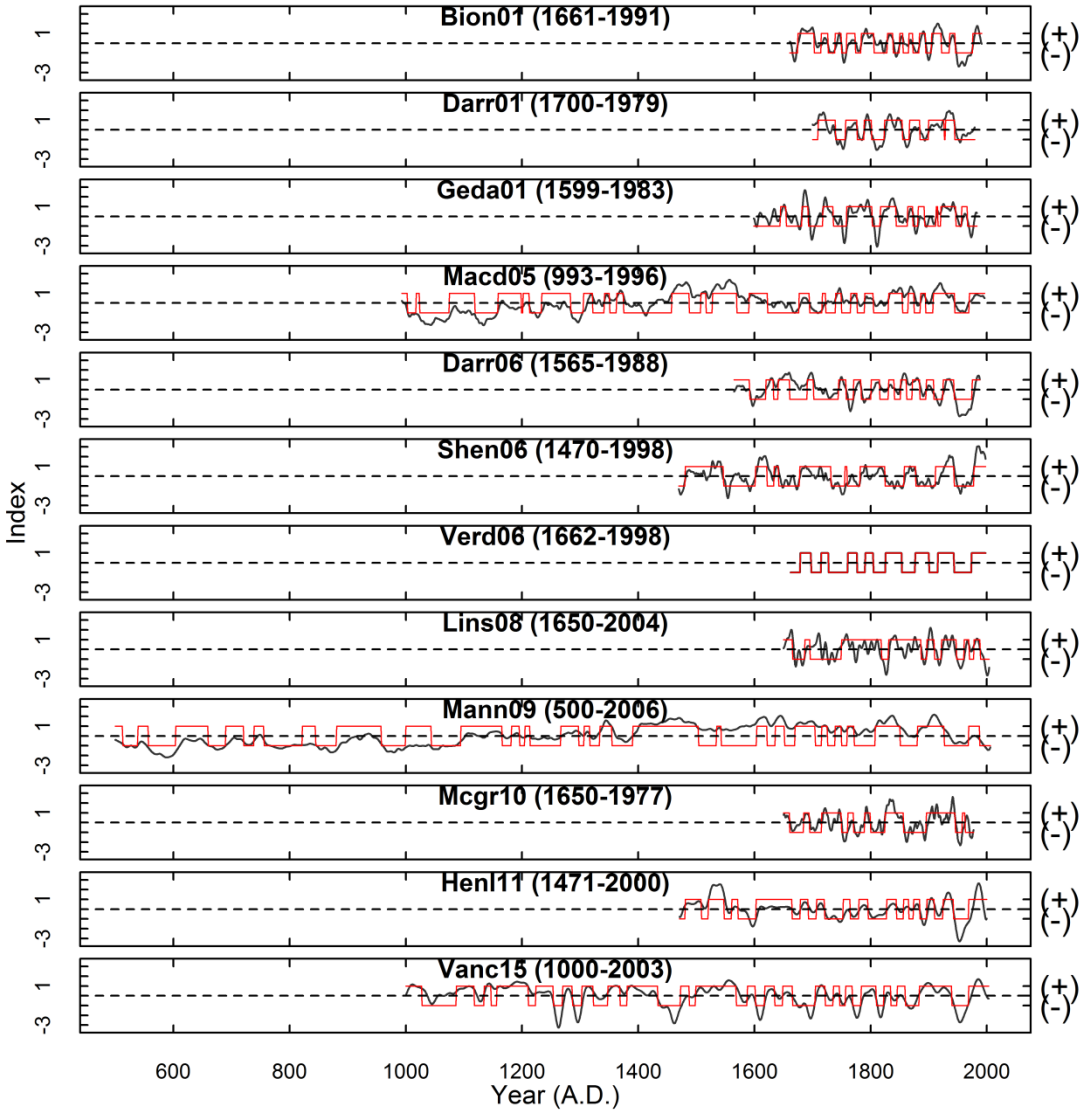


Figure 6 Density plot of run length samples by applying different window widths to all reconstructions, (a1-a3): comparison between window width 20 and 40, in which (a1) is unconditional run length distribution, panel (a2) is the distribution conditioned on run lengths less than 20 years and panel (a3) is the distribution conditioned on run lengths greater than 20 years; (b1-b3): comparison between window width 20 and 60, in which (b1) is unconditional run length distribution, panel (b2) is the distribution conditioned on run lengths less than 30 years and panel (b3) is the distribution conditioned on run lengths greater than 30 years



1

2 **Figure 7 Filtered PDV reconstruction time series (black line) with extracted run length (red line)**

3

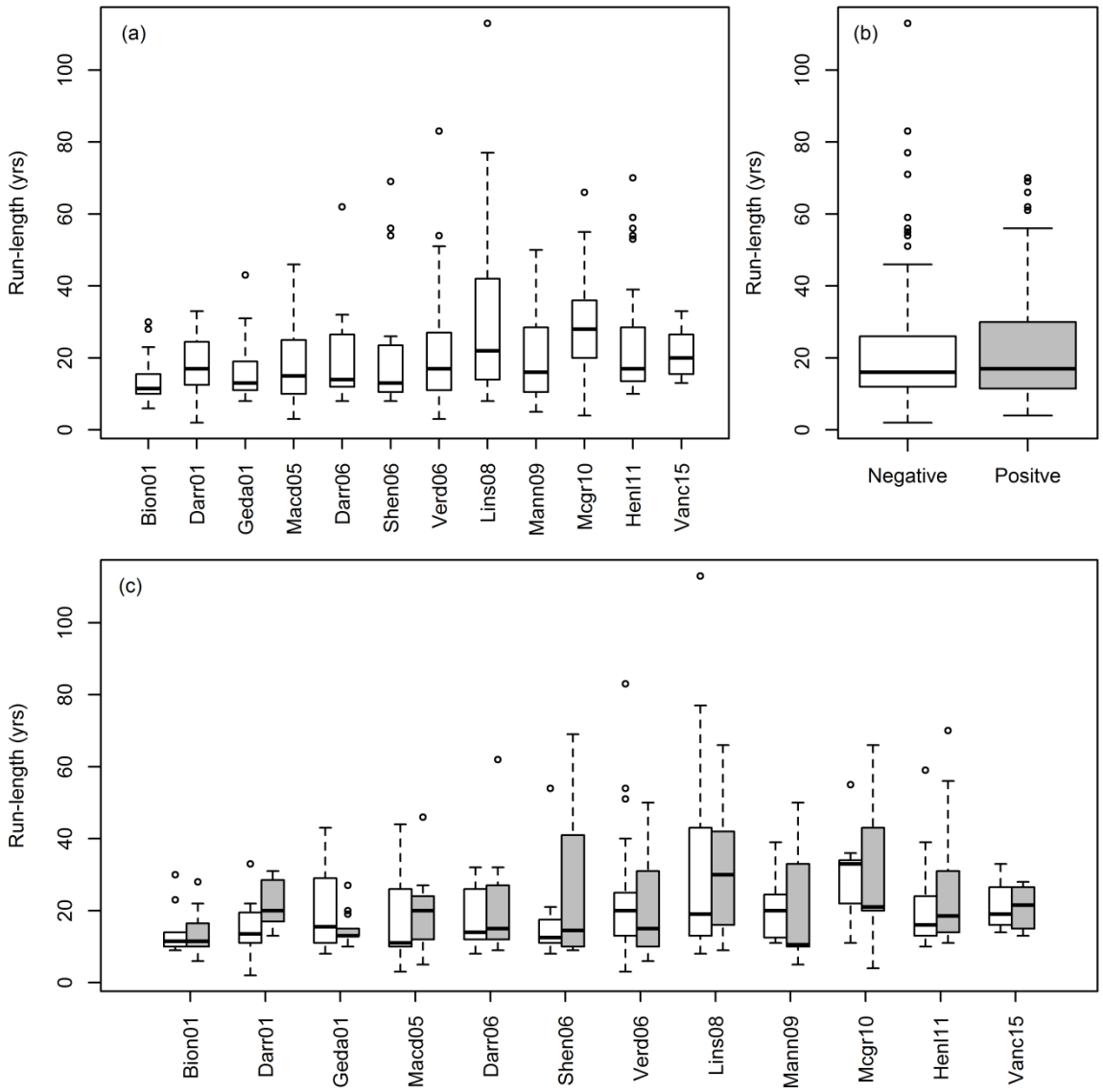


Figure 8 Boxplot of run lengths samples from different reconstructions: (a) all run lengths for each reconstruction; (b) pooled positive (grey) and negative (white) run lengths; (c) positive (grey) and negative (white) run lengths for each reconstruction

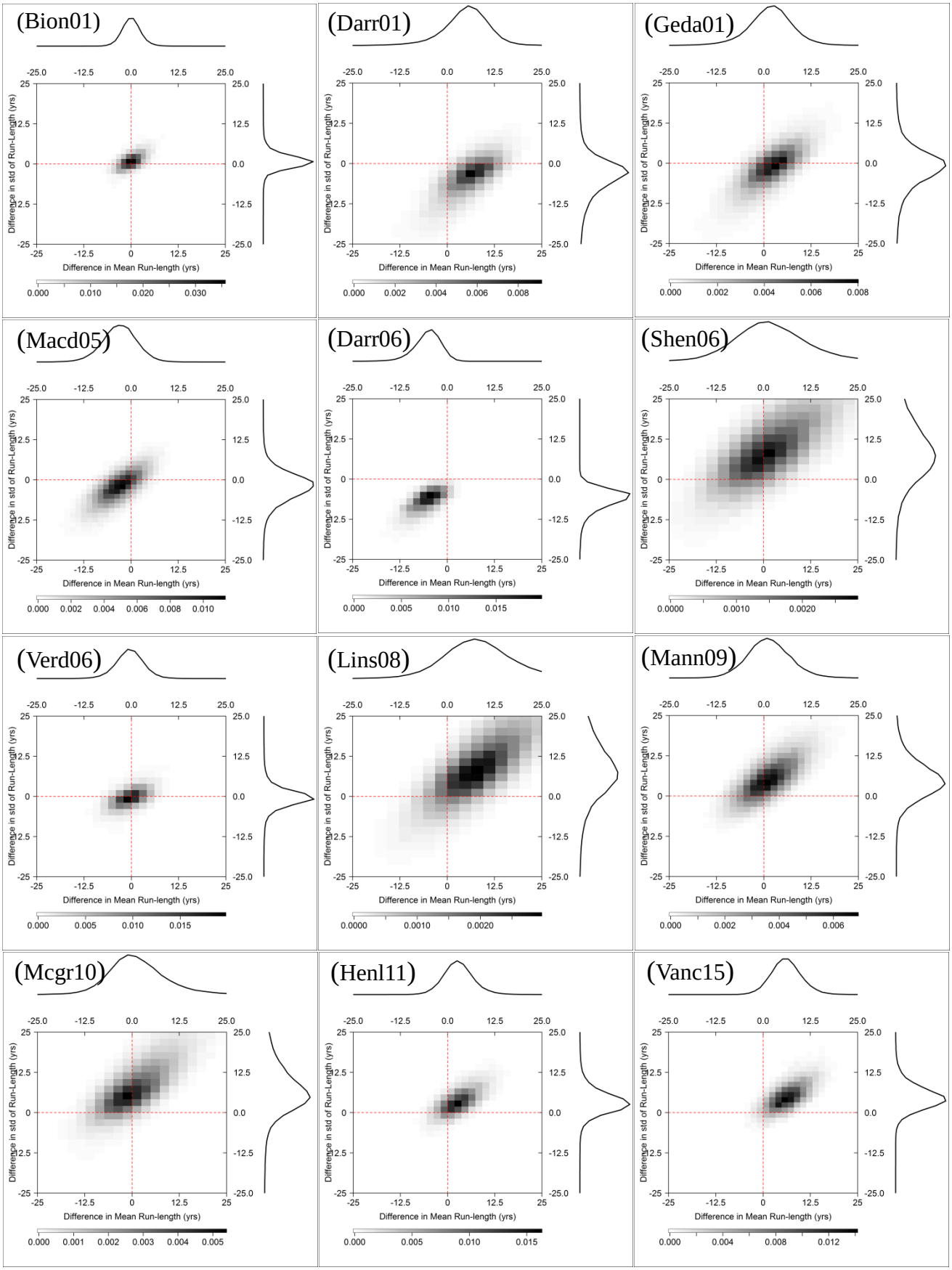


Figure 9 Density and scatter plots of the differences between mean and standard deviation of positive and negative run length simulations (positive minus negative)

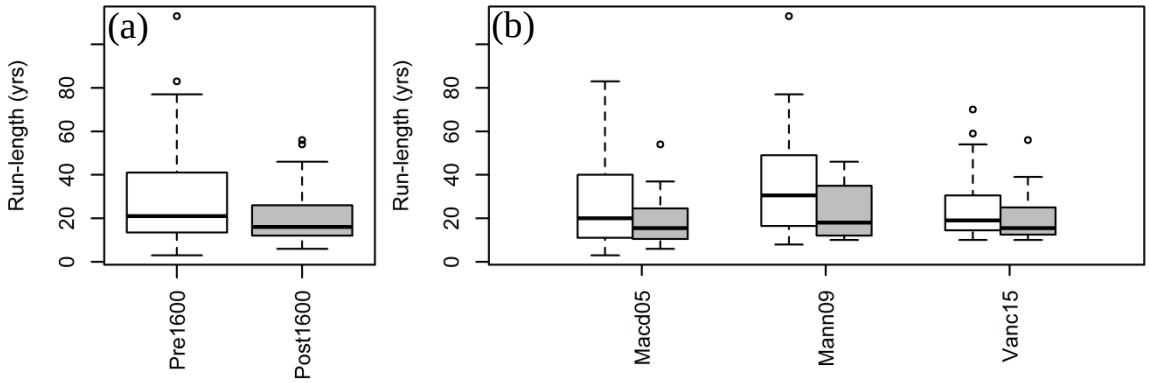


Figure 10 Boxplot of run lengths samples from different reconstructions (white box indicates pre-1600 and grey box indicates post-1600): (a) pooled run lengths from all three reconstructions; (b) run lengths for each reconstruction

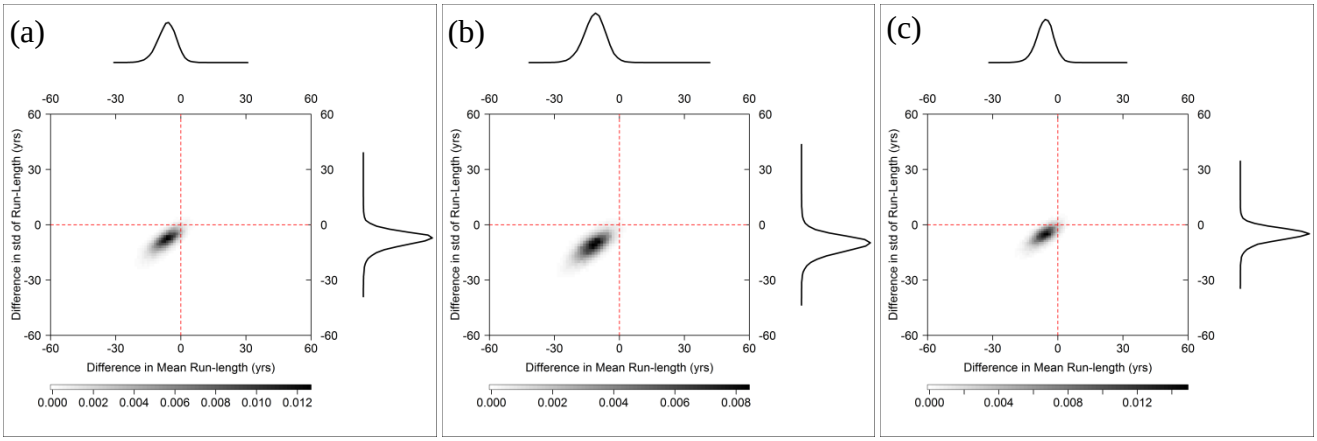


Figure 11 Density and scatter plots of the differences between mean and standard deviation of pre-1600 and post-1600 run length simulations (post-1600 minus pre-1600): (a) Macd05; (b) Mann09; (c) Vanc15