Detecting marine heatwaves with sub-optimal data

Robert W. Schlegel1,2,\*, Eric C. J. Oliver1, Alistair J. Hobday3, Albertus J. Smit2

1Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada  
2Department of Biodiversity and Conservation Biology, University of the Western Cape, Bellville, South Africa  
3CSIRO Oceans and Atmosphere, Hobart, Tasmania, 7000, Australia

\_\_\*Correspondence:\_\_  
Robert Schlegel  
[robert.schlegel@dal.ca](mailto:robert.schlegel@dal.ca)

**Keywords: marine heatwaves, sea surface temperature, sub-optimal data, time series length, missing data, decadal trend**

# Abstract

Marine heatwaves (MHWs), or prolonged periods of anomalously warm sea water temperature, have been increasing in duration and intensity globally for decades. However, there are many coastal, oceanic, polar, and sub-surface regions where our ability to detect MHWs is uncertain due to the unavailability of high quality data. Here we investigate the effect that short time series length, missing data, or linear decadal temperature trends may have on the detection of MHWs. We show that MHWs detected in time series as short as 10 years did not have durations or intensities appreciably different from events detected in a standard 30 year long time series. We also show that the output of the MHW algorithm for time series missing less than 20% data did not differ appreciably, and that this could be increased to 40% when gaps were filled with linear interpolation. Linear decadal trends of 0.10°C/dec or greater added to a time series could cause worryingly large increases in the count and duration of detected MHWs, but this could be controlled for with reasonable confidence. The decadal trend in a time series has the largest effect on the detection of MHWs, but is also the easiest to correct for. Time series length has less of an effect on MHW detection than missing data, but is the most difficult to correct for. We provide suggestions for best practices to improve the accuracy of MHW detection with sub-optimal time series and show how the accuracy of these corrections may change regionally around the world.

# Introduction

The idea of locally warm seawater being problematic is not a novel concept. We have known for decades that seemingly transient warm water occurrences in the ocean could result in major impacts (e.g. Baumgartner, 1992; Salinger et al., 2016). The study of the effects of anomalously warm seawater temperatures began in earnest in the early 1980’s when research into the ENSO phenomenon began (e.g. Philander, 1983). After the 1980’s, researchers began noticing that warm water events were becoming more frequent and problematic, but it wasn’t until 2018 that this was demonstrated with global observations (Oliver et al., 2018).

In order to quantify the increased occurrence and severity of these events it was necessary to develop a methodology that would be inter-comparable for the entire planet. This was accomplished in 2016 after the International Working Group on Marine Heatwaves (marineheatwaves.org) initiated a series of workshops to address this issue (Hobday et al., 2016). This definition for anomalously warm seawater events, known as marine heatwaves (MHWs), has seen wide-spread and rapid adoption due to ease of use and global applicability. One problem with this algorithm that has not yet been addressed is the assumption that a researcher has access to the highest quality data available when detecting MHWs. In the context of MHW detection, a ‘high quality’ time series is spatio/temporally consistent, quality controlled, and at least 30 years in length (Table 3 in Hobday et al., 2016). While not stated explicitly in Hobday et al. (2016), a ‘high quality’ time series should also not have any missing days of data. To avoid contention on the use of the word ‘quality’, time series that meet the aforementioned standards are referred to here as ‘optimal’, whereas those that do not meet one or more of the standards are referred to as ‘sub-optimal’. Another unresolved issue with the Hobday et al. (2016) algorithm, which does not fall within the requirements for optimal data, is how much of an effect the long-term (secular) trend in a time series may have on the accurate detection of MHWs.

Most remotely-sensed data, and more recently output from ocean models and reanalyses, consist of over 30 years of data and utilise statistical techniques to fill gaps in their time series from a number of environmental and technical sources. This means that these data are considered optimal for MHW detection. An excellent reference for the remotely-sensed products currently available, as well as their strengths & weaknesses, is Table 12.3 in Harrison et al. (2019). Even though remotely-sensed data products are considered optimal, they still have other issues (e.g. land bleed and incorrect data flagging) and so it may be recommended that for coastal applications researchers utilise sub-optimal data, such as sporadically collected *in situ* time series (Hobday et al., 2016; Smit et al., 2013).

This paper seeks to understand the limitations the use of sub-optimal data impose on the accurate detection of MHWs. Of primary interest are three key challenges:

1. The use of time series shorter than 30 years
2. The use of temporally inconsistent (missing data) time series
3. The use of time series in areas with large long-term temperature trends

We will use a combination of reference time series from specific locations and global data to address these issues. The effects of the three sub-optimal data challenges on the detection of MHWs are quantified in order to provide researchers with the level of confidence they may express in their results. Where possible, best practices for the correction of these issues are detailed.

# Defining marine heatwaves

The definition used in this paper for a MHW is “a prolonged discrete anomalously warm water event that can be described by its duration, intensity, rate of evolution, and spatial extent.” (Hobday et al., 2016). This qualitative definition is quantified with an algorithm that calculates a suite of metrics. These metrics may then be used to characterise MHWs and to effectively compare them against known ecological/financial impacts. The calculation of these metrics is made possible by first determining the mean and 90th percentile temperature for each of the 366 calendar day-of-year (doy) in a time series. The mean doy temperatures, which also represent the seasonal signal in the time series, provide the expected baseline temperature whose daily exceedance is used to calculate the intensity of MHWs. The 90th percentile doy temperatures serve as the threshold that must be exceeded for 5 or more consecutive days for the anomalously warm temperatures to be classified as a MHW and for the calculation of the MHW metrics to begin.

In this paper we will focus on the two metrics that most succinctly summarise a MHW. The first metric, *duration (days)*, is defined as the number of days that the temperature remains at or above the 90th percentile threshold without dipping below it for more than 2 consecutive days. The duration of an event may be used as a measurement of the chronic stress that a MHW may inflict upon a target species or ecosystem [e.g. citations needed]. The second metric, *maximum intensity (°C)*, is the single warmest day during the event and is calculated by subtracting the mean doy temperature on that day from the recorded temperature. This metric may be used as a measurement of acute stress [e.g. citations needed]. A third metric, *cumulative intensity (°C)*, will also be used in this paper to determine the ‘largest’ MHW in a time series (see Methods). This metric represents the sum of temperature anomalies over the duration of the MHW and is comparable to the degree heating days metric used in coral reef studies [citation needed]. There are many other MHW metrics and the full explanation for them may be found in Table 2 of Hobday et al. (2016).

The Hobday et al. (2016) MHW definition has been developed for python (<https://github.com/ecjoliver/marineHeatWaves>), R (Schlegel and Smit, 2018), and MATLAB (Zhao and Marin, 2019). Before beginning the analysis proposed here, we compared the R and python default outputs, assessed how changing the arguments affected the results, and compared the other functionality provided between the two languages. While some style differences exist between the added functionality of the languages, the core climatology outputs are identical to within < 0.001 °C per day-of-year (doy). An independent analysis of the Python and MATLAB results also confirmed that they were functionally identical (pers. com. Zijie Zhao; MATLAB distribution author).

# What are optimal data for detecting marine heatwaves?

When working with outliers in a time series, such as MHWs, it is important that the quality of the data are high [citation needed]. Hobday et al. (2016) stated that high quality data, referred to here as ‘optimal’, used for the detection of MHWs should meet the following criteria:

1. A time series length of at least 30 years
2. Quality controlled
3. Spatially and temporally consistent
4. Be of the highest spatial and temporal resolution possible/available
5. *In situ* data should be used to compliment remotely sensed data where possible

Whereas the authors did not specifically state that time series must not contain large proportions of missing data, it can be inferred from the aforementioned requirements. Another issue affecting the accurate detection of MHWs not discussed in Hobday et al. (2016) is the presence of long-term trends in a time series. Oliver et al. (2018) have shown how dominant the climate change signal can be in the detection of MHWs and we seek to quantify this effect here.

A time series with a sub-optimal length may impact the detection of MHWs by negatively affecting the creation of the daily climatology relative to which MHWs are detected in two primary ways. The first is that with fewer years of data to draw from, the presence of an anomalously warm or cold year will have a larger effect on the climatology than with a sample size of 30 years. The second cause is that because the world is generally warming (Pachauri et al., 2014), the use of a shorter time series will almost certainly introduce a warm bias into the results.

The climatology derived from a time series serves two main roles (Organization, 2017); 1) it serves as a ‘benchmark’ relative to which past and future measurements can be compared, and against which anomalies can be calculated, 2) it reflects the typical conditions likely to be experienced at a particular place at a particular time. The WMO Guide to Climatological Practices (Organization, 2011) stipulate that daily climatologies (which they call ‘climate normals’) must be based on the most recent 30-year period that ends on a complete decade (currently 1981 – 2010). It is from this WMO guideline that the optimal length of time series for MHW detection was derived.

Some remotely sensed products suffer from ‘gappiness’ that result in missing data being introduced. This may be due to cloud cover, the presence of sea ice, unsuitable sea states, etc., which become more prevalent at smaller scales, particularly nearer the coast. Some products smooth out these influences, but this results in smoothed sea surface temperature (SST) fields that may mask small-scale spatial variations in surface temperatures. Remotely sensed products may also fill gaps by blending with data from other products, which may have its own suite of consequences. It is also known that coastal SST pixels from remotely-sensed products may have biases in excess of 5°C from *in situ* collected data (Smit et al., 2013). This is why the use of imperfect *in situ* collected time series may still be encouraged for coastal applications. These data are however also prone to large gaps and so the problems these data face with regards to accurate event detection are also uncertain. It must be stressed here that the best practices for working with sub-optimal data presented in the Discussion section do not address the issues that remotely sensed data have near coastlines.

# Methods

To quantify the effects that time series length, missing data, and long-term trends have on MHW detection we will be comparing the count, duration (days), and max intensity (°C) of MHWs from time series as they become increasingly sub-optimal. To allow for roughly equal sample sizes across all tests, only the results for MHWs in the final 10 years of data (2009 – 2018) are used. These comparisons will be made for the averages of MHWs in the final ten years of each test, as well as the single largest MHW. This will provide us with more insight into the broad effect that sub-optimal data have on MHW detection as well as allowing us to look at the specific effect on a single event without inundating us with results.

<!-- - These are a more qualitative result that may be more applicable to a broader audience. -->

We then use the following three questions to frame the results:

1. How sub-optimal may a time series become before any of the above items are significantly different from those calculated with an optimal time series?
   * For example, how short may a time series be before the mean duration of MHWs in it differ from the full 30 year time series?
2. What amounts of uncertainty are introduced into the results from the increasingly sub-optimal data?
   * For example, when 20% of a time series is missing how much might the max intensity of a single MHW be effected compared to that same MHW when detected in the same time series with no missing data?
3. Are the uncertainties introduced by sub-optimal data the same/similar everywhere in the world, or do they differ based on some observable pattern/known oceanographic feature(s)?
   * For example, when a time series has a long-term trend of 0.2°C/dec in an eastern boundary upwelling system (EBUS), does the effect this have on the maximum intensity of the events differ from the same trend in a time series in a western boundary current (WBC)?

To answer these three questions we will use the remotely sensed NOAA OISST dataset (Reynolds et al., 2007, @Banzon2016). This daily remotely-sensed global SST product has a 1/4 degree spatial resolution with the first full year of sampling being 1982. These data are interpolated where possible against a database of *in situ* collected temperatures so that the final product does not have any spatial or temporal gaps. Before being used in this study each time series (pixel) is de-trended by fitting a simple linear model and removing the residuals from the data. This must be performed so that we may control for the effects of time series length and decadal trend separately. Once de-trended, each time series is treated to the suite of sub-optimal controls (see following sub-sections) and the desired results itemised above are extracted. In order to ensure as much consistency as possible across all test results, the entire duration of the time series will be used as the climatology base period, and not the 30 year standard recommended by the WMO. See Appendix 1 for the difference between the focus results when using a 30 year base period vs. a 37 year base period.

Significant differences in MHWs from the time series with different sub-optimal tests will be found by comparing them against the same optimal reference time series (i.e. 30 year length, no missing data, no long-term trend) with a pairwise post-hoc Kruskal-Wallis test (Siegel and NJ, 1988). This non-parametric test looks for differences between the central tendency of a control and a series of comparison sets of data, providing the probability (*p*-value) that the sets of data have been drawn from the same pool. It was necessary to use a non-parametric test because we cannot ensure that the data being compared are always normally distributed or have a similar count of observations. We think that the inclusion of a *p*-value in this way is important to frame the results, but are not using an arbitrary *p*-value threshold to reject a null hypothesis that the outputs from the sub-optimal data are different from the optimal data because testing for null hypotheses in this way is becoming increasingly discouraged (Wasserstein et al., 2019). That being said, we will still highlight comparisons that generate a *p*-value of 0.05 or less.

The amount of uncertainty that the sub-optimal tests introduce into the results will be calculated by measuring the percent of change in the results from the control (optimal) time series as the data become more sub-optimal. This measurement will be performed on the mean of the MHWs detected in the final 10 years of data and the single largest MHW. No significance test is used here, rather the range of change in these values is shown in order to provide a benchmark against which one may decide how much change is too much.

With the changes in MHW detection due to sub-optimal data now calculated, linear models will be used to quantify the rates of change that these tests introduce. These trends will be visualised on a global scale to provide us with insight into any spatial patterns that may exist.

The answers to the first two questions posed above will be highlighted with the three reference OISST time series from Hobday et al. (2016). These time series are taken from the coast of Western Australia (WA; Figure 1A), the Northwest Atlantic Ocean (NWA; Figure 1B), and the Mediterranean Sea (Med; Figure 1C). These time series are used here for ease of reproducibility and because they each contain a MHW that has been the focus of multiple publications. These results will be overlaid on top of 100 randomly selected pixels from the global OISST dataset to show a more thorough range of results.

It should be noted that while not a specific focus in this study, the effects that the sub-optimal tests have on the seasonal mean and threshold climatologies may also be of interested and so have been included in Appendix 2.

The following three sub-sections describe how the three sub-optimal time series tests will be controlled for.

## Controlling for time series length

There are currently 37 complete years of data available in the NOAA OISST dataset (1982 – 2018). In order to determine the effect that time series length has on the focus output we will systematically shorten each time series one year at a time from 37 years down to 10 years (2009 – 2018), before running the MHW detection algorithm. The focus outputs for each one year step for each of the time series will then be compared against the output from the optimal 30 year version of that same time series as outlined above.

In order to ensure equitable sample sizes we will only be comparing the MHW metrics for events detected within the last 10 years of each test as this is the period of time during which all of the different tests overlap. This is also why we have limited the shortening of the time series lengths to 10 years, so that we could still have a reasonable sample size to draw from.

Because the lengths of the time series were being varied, and were usually less than 30 years in length, it was also necessary that the climatology periods vary likewise. To maintain as much consistency as possible across the results we used the full range of years within each shortened time series to determine the climatology. For example, if the time series had been shortened from 37 to 32 years (1987 – 2018), the 32 year period was used to create the climatology. If the shortened time series was 15 years long (2004 – 2018), this base period was used. The control time series were those with a 30 year length ending in the most recent full year of data available (1989 – 2018). Note that due to necessity this differs from the climatology period of 1982 – 2011 that would mostly closely match the WMO standard. The effect of shifting the 30 year climatology base is discussed in Appendix 1.

The *a-priori* fix proposed to address the issue of short time series length is to use a different climatology estimation technique. The option currently available within the MHW detection algorithm is to expand the window half width used when smoothing the climatology. Other techniques, such as harmonic regression/Fourier analysis, would have a similar effect but are not used here in favour of the Hobday et al. (2016) methodology. It is beyond the scope of this paper to compare every possible climatology calculation technique.

## Controlling for missing data

In order to determine the effect of random missing data on the focus outputs, each time series will have 0 – 50% of its data removed in 1% steps before running the MHW algorithm. The control time series is that missing 0% data.

The *a-priori* fix for the issue of missing data in the time series is to linearly interpolate over any gaps. There are many methods of interpolating (imputing) gaps in time series, such as spline interpolation, but we choose linear interpolation here due to its speed, simplicity, and the lack of assumptions it imposes on the data. It is beyond the scope of this paper to account for every possible method of interpolation however; we will also count the lengths of the gaps over which data are interpolated in order to show at how many consecutive missing days interpolation begins to fail.

## Controlling for long-term trends

To quantify the effect of long-term (secular) trends on the focus outputs we will add linear decadal trends of 0.00 – 0.30°C/dec in 0.01°C steps to each time series. The difference this causes in the outputs will be quantified with the same tests as for length and missing data. The optimal time series used for control will be those with no added trend (e.g. 0.00°C/dec).

There is no proposed *a-priori* method to correct for the added linear decadal trend in these data as this would be simply not to add a trend. Rather it is proposed that the relationship between the slope of the added trend and the results it has on the focus outputs be documented to determine if a predictable relationship may be used to correct the results *post-hoc*.

# Results

## Time series length

Shortening the length of a time series from 30 to 10 years had an unpredictable effect on the count of MHW detected in the most recent 10 years of data (Figure 2A). We see that at a 10 year length, 90% of the pixels we tested had somewhere between 10 more to 9 fewer MHWs than the 30 year control. The increase or decrease in the count of MHWs for each time series was somewhat linear, meaning that one may be able to say what the change in the count of MHWs may be, but it does not allow us to say if this change will be positive or negative. In Figure 2D we see that the change in the sum of the MHW days for a 10 year time series ranges primarily from 49% fewer to 59% more than the 30 year control. This change is slightly more linear than for the count of MHWs, but again we are faced with the problem that values may increase or decrease. The mean of the maximum intensities of the sample MHWs are also seen to either increase or decrease at roughly similar rates, with 10 year time series having mean intensities anywhere from 13% less to 5% more than the 30 year control. There is no point where any of these samples become significantly different (*p*-value <= 0.05) than the control sample.

It is important to note that increasing the climatology period longer than 30 years had almost as rapid an effect on creating dissimilar outputs as using fewer years of data does. This was an unexpected result that stresses the importance of adhering to the WMO standard as closely as possible to ensure the comparability of results.

Looking at the effect that shortening time series length has on the duration and maximum intensity of the largest MHW from each time series we see that both values tend to decrease (Figure 3DG), while the count of MHWs within the duration of the focus MHW increase (Figure3A). This is because the shortening of a time series may increase the seasonal or threshold climatologies, so the shorter a time series becomes, the lower the maximum intensity and shorter the duration of the MHWs may become. MHWs with many spikes, rather than a smooth hump, will be particularly effected by this change in the climatology as it will raise up enough to break the MHW into many smaller pieces, as shown by the increase in MHW count in Figure 3A.

We see in Figure 4 that there are clear global patterns in the changes in MHW results as time series are shortened from 30 to 10 years. The median change in the count of MHWs due to changes in time series length is nearly 0, but much of the western Pacific and northern Atlantic oceans show huge rates of increasing MHW counts as time series are shortened (Figure 4A). The rates of change in the eastern Pacific, southern Atlantic, and the Indian Ocean show a mix of both increasing and decreasing counts of MHWs as time series become shorter. The patterns of change in the sum of MHW days very closely resembles the change in the count of MHWs (Figure 4B). The decreasing maximum intensities of MHWs caused by decreasing time series lengths were also generally seen to occur throughout the oceans (Figure 4C). The median rate of change caused by decreasing time series length from 30 to 10 years is seen to be 0.21% per year. This means that, on average, a MHW detected in a 10 year time series will have a maximum intensity about 8.4% cooler than a MHW detected in a 30 year time series (0.21%/year times 20 year difference). This is a very small margin and shows the robustness of the MHW detection methodology. There are however areas of the oceans where decreasing a time series tends to increase the maximum intensities of the MHWs detected. These areas are roughly the same parts of the oceans where the shortening of a time series causes a decrease in the sum of MHW days detected. It is important to note that the long-term trends in these data were removed beforehand so the patterns observed in Figure 4 are due to the properties of the time series themselves and not the climate change signal that would otherwise be dominant in the results.

RWS: Figure 4DEF show the results of the time series length on the largest MHW in the time series but I didn’t go over the results here as I felt the results section was getting too long… It may be that they don’t need to be included.

## Missing data

The effects of increasing missing data on MHW detection tended to be more linear than for time series length. Up to 25% missing data, the count of MHWs in a times series may increase by 10 or decrease by 11 (Figure 2B). Past this point the count of MHWs falls at a roughly linear rate until there are 3 – 22 fewer MHWs per time series when 50% data are missing. The effect of missing data on the sum of the MHW days was surprisingly linear at a rate of roughly 2% fewer MHW days in a time series for every 1% of missing data (Figure 2E). The effect of missing data on the maximum intensities of the MHWs was also linear, but very noisey. The average maximum intensities of MHWs detected in time series missing 50% of their data could increase by 3%, or decrease by 34%.

The effect of random missing data on the single largest MHW in each time series was dramatic. As missing data in a time series increase it becomes increasingly likely that the largest MHW in a time series will be dissected into multiple smaller events. In Figure 3B we may see that it is not uncommon for this dissecting to begin with as little as 1% missing data, and increases in severity up to 25 – 30%. From this point the number of separate events the MHW is dissected into decreases as even the smaller events dissected off of the whole are obliterated by the loss of data. The duration of the focus MHW was almost always negatively impacted by missing data (Figure 3E). Even though the decrease in duration due to missing data is very rough, we see that it follows a linear trend of a reduction ranging from 1 – 3% per 1% of missing data. At 30% missing data it became common that the largest MHW would be removed entirely from the time series.

The effect that missing data can have on a MHW depends largely on the shape of the MHW, which is the area above the threshold climatology and below the observed anomaly. The WA event has a very pronounced peak (Figure 1A), so when more data are missing we see how likely it becomes that this peak is not being recorded. The maximum intensity measured in the control time series is 6.5°C, but we see that because very few days of this MHW were so intense, increases in missing data become more likely to remove these large values and the maximum intensity of the WA event begins to decrease more rapidly than either the NWA or Med MHWs.

The global patterns in missing data are unremarkable and generally consistent across the oceans (Appendix 3A).

## Long-term trends

The effect added decadal trends had on MHW detection were unsurprisingly the most linear of the three tests and resulted in the largest changes in the results. Very rarely an added decadal trend would lead to a reduction in the count of MHWs in a time series, but generally it caused a linear increase at roughly 8 additional MHWs detected for every 0.1°C/dec added. The effect that all of these additional MHWs had on the sum of the MHW days was an increasing ranging from 18 – 90% for every 0.1°C/dec added (Figure 2F). The effect of decadal trends on the maximum intensity of MHWs, though generally linear, could be either positive or negative at a rate of -9 – 7% per 0.1°C/dec added.

The largest MHW in each time series was never dissected into multiple events due to the trend in the data, but at an added trend of 0.19°C/dec we began to see MHW disappearing from the record (-1 for count; Figure 3C). This is caused by the threshold climatology raising up enough due to the added decadal trend that the temperature anomaly sinks below the surface. The duration of the MHWs were affected differently by the added linear trend. The Med showed practically no increase in duration due to decadal trend, the WA event saw a large jump at 0.03°C/dec, and the NWA had a dramatic jump at an added trend of 0.09°C/dec, followed by a few other increases at higher added trends (Figure 3F). Likewise we may see that all of the other 100 pixels included in Figure 3 tend to jump up or down in dramatic steps. This happens because the temperature anomalies are increasing in magnitude more rapidly than the threshold and in some instances nearby MHWs will raise up enough that they connect to their neighbours. Creating a sort of MHW land bridge. The effect that the linear trend had on the maximum intensity of each event was a simple linear function of the decadal trend and where in the time series the event occurred (Figure 3I). As in Figure 3C, at an added trend of 0.19°C/dec we see that some of the focus MHWs sink below the raising threshold and the change in maximum intensity drops to a value of -100% (Figure 3I).

The global patterns in added decadal trends generally show that MHW metrics also increase (Appendix 3B).

## Potential fixes

RWS: I’ll insert a table in this section that shows the slopes in the rates of change for the results. I’ll also include examples of how one may perform fixes on sub-optimal results using examples. This may require that I create an additional figure to help illustrate the process.

The *a priori* fix proposed for shorter time series, that of creating a smoother seasonal signal, created results with more similar overall days of MHWs per year, but all of the focus results were less similar than without applying the fix (results not shown; RWS: Could include as an Appendix if desired). This happened because the over smoothing of the seasonal signal prevents the detection of the smaller MHWs in the time series while inflating the size of the larger MHWs. A larger issue caused by short time series is the amount that the centre of the climatology increases or decreases, more so than the increase in variability caused, which additional smoothing seeks to address. This is not something that can be controlled for *a-priori* and is better controlled for in a *post-hoc* manner along the same lines as the proposed fix for decadal trends (see below).

Linear interpolation was proposed as an *a priori* fix to address the issue of missing data and was very effective. This fix could potentially allow for the use of time series missing up to 40% of their data (Figure 4), assuming that there is not so much missing data that the period of time during a MHW that one may be wanting to study/isolate is completely missing.

RWS: I’ll write here about consecutive missing days of data and how that interferes with linear interpolation. This will be shown as Figure 5.

Given the relationship between decadal trends and the focus outputs seen above, one may be able to apply a *post-hoc* correction to one’s results if the decadal trend in a given area is known.

RWS: I still need to flesh out the corrections for decadal trends. I do not plan on creating a figure for this.

RWS: It must also be shown how taken with an area of occurrence the fixes can actually be zeroed in on at a great rate.

# Discussion

An investigation into the effects that sub-optimal data have on MHWs revealed that there are no firm statistical thresholds within which the outputs of the MHW detection algorithm become different than results generated by optimal data. The ranges of the size of effects that sub-optimal data have on MHW results was able to be determined and it is up to the user to decide for themselves what an acceptable uncertainty may be. We discuss here what these ranges of uncertainty may mean for MHW research.

The MHW results from 10 years of data are not appreciably different from those detected with 30 years that they overlap with (i.e. the final 10 years of data). The rates at which the count, duration and intensity of events change from year-to-year within a single time series may change wildly, but a global sampling showed that overall there is a linearly increasing range in the possible change in results one may expect. The rates seen in Table 1 may therefore be applied *post\_hoc* to MHWs detected in shorter time series in order to make the results comparable to those from an optimal time series.

An unexpected result was that increasing the length of a time series longer than 30 years reduced the probability that the outputs would be comparable by as much as shortening the time series did. This means that the common (often unspoken) assumption that using 30 years of data is the same as using > 30 years of data is incorrect. In other words, the 30 year length is often thought of as a minimum length needed to constrain the climatology but we have shown here that using a climatology period greater than 30 years may create outputs as different as using fewer than 30 years. This is due to the interannual variability in a time series. In more stable time series with dependable seasonal signals there will be no appreciable difference between results calculated with a 30 year base period versus the full 37 year base period. In a very chaotic time series with an unreliable seasonal signal, such as the WA time series, increasing the base period is only increasing the overall variance that the creation of the seasonal mean must contend with. It is therefore important to stress the adherence to the WMO standards for climatology periods as closely as possible, should one want one’s results to be comparable to other studies. Increased smoothing of the climatologies derived from shortened time series was not an effective fix, in fact it made the results worse and we strongly recommend that the default climatology creation methodology in Hobday et al. (2016) be adhered to.

The MHW algorithm proved to be resilient to missing data as long as there are not gaps larger than five days. Time series missing as much as 30% of their data may be used if necessary as the effect this may have on the count of MHWs is comparable to using a 10 year time series. Over 30% missing data and the count of MHWs in a time series is effected too much for the results to be reliable. That being said, the effect that missing data has on the overall MHW days in a time series is very predictable and can be corrected for rather easily as seen in Table 1. We showed here that a simple correction for missing data in a time series is to linearly interpolate over the gaps. It is not however recommended to do this with more than 40% missing data as this begins to dramatically distort the MHW detection algorithms ability to compute metrics for individual MHWs. If this is necessary to do for some reason, the resultant MHWs for the entire time series may still be used to infer the chronic and acute stress that organisms may face in a given location, but any individual events detected should not be taken as an accurate recording.

The decadal trends in times series have the largest potential effect on the MHWs detected. These effects are however the most predictable of the three tests and when taken with time series length and the year in which an event in question has occurred it is possible to infer a correction for the maximum intensity. The effect this has on the duration can also be worked out by considering the general rise (or fall) in the seasonal mean and threshold climatologies and how that may engulf neighbouring days or even other events. A concept to consider with the increase in duration from added decadal trends is that the temperatures in the time series increase “faster” than the 90th percentile threshold. So as the decadal trend increases, a given MHW effectively spreads outwards. If the rate of onset/decline for the MHW was more gradual (e.g. the NWA event) it will increase in duration more rapidly. If the rate of onset/decline was more rapid (e.g. the Med event), then the duration of the MHW won’t change much with a larger decadal trend. If MHWs have close neighbouring MHWs then as they spread outward they may encounter each other and merge into a single event. This reduces the overall count of the MHWs detected in a time series while increasing the mean duration of the events detected.

RWS: An issue that needs to be discussed (according to a reviewer) is how increased heat content of sub-surface waters makes MHWs below the surface appear to last longer or be larger etc.

RWS: Perhaps a paragraph that talks about why using p-values for determining an acceptable threshold may be bad/problematic. I’ve not written this as it turned out that p-values were a complete dead end and aren’t used anywhere in the results or discussion.

# Conclusions

The acceptable sub-optimal data limits, the amount of uncertainty they introduce into the results, and their proposed corrections are as follows:

1. Time series length:

* A length of 10 years produces acceptable MHW metrics that may be used with some caution
* The uncertainty introduced into results by time series length is roughly +-xxx/year
* Smoothing the climatology before detecting MHWs makes the results **worse**, not better, and should not be done

1. Missing data:

* Missing data up to 30% produces results as dissimilar from a control time series as using 10 years of data does
* The detected metrics for single large MHWs are no longer reliable with more than 20% missing data
* Linear interpolation is an excellent fix for missing data up to 40%, assuming one does not have periods of consecutive missing data in excess of five days

1. Long-term trends

* Long-term trends had a greater effect on MHWs than the other two sub-optimal tests, but can be corrected for with the most confidence
* If one knows the decadal trend present in a study area it is possible to correct for the max intensity of a MHW by subtracting the year of occurrence as a function of the slope present in the data
* For example, if the decadal trend in an area is 0.30°C/dec, and the peak date of an event is in the 25th year of a 30 year time series (83% of the length) then the the impact of the decadal trend on the max intensity will be roughly: 0.3\*0.83 = 0.25(°C)

We have shown here that researchers must not shy away from the use of sub-optimal time series when the situation calls for it, such as coastal research or sub-surface analyses. Time series length may have an unpredictable effect on MHW results, but this may be corrected for within reason, and we have shown that time series lengths as short as 10 years are still useful for MHW research. Any shorter than 10 years however and the relationship between time series length and the effect on MHW metrics becomes too unpredictable to provide any corrections with confidence. Missing data has a larger effect on MHW detection, but is less of a concern as linear interpolation can largely fix the challenges this creates, up to a threshold of 40% missing data. Lastly, the effect of long-term trends on MHW detection are the most predictable and when taken with time series length may be corrected for with a reasonable degree of certainty. The MHW detection algorithm is very robust and we have shown here that one may be confident in the inter-comparability of one’s results when using time series within a generous range of sub-optimal data challenges.

# Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Author Contributions

The majority of the text and figures were produced by RWS. A large portion of an early version of the text and a number of initial figures were produced by AJS. AJH, ECJO, and AJS provided several rounds of comments on the manuscript as it was developed. RWS synthesised the comments and uploaded the manuscript.

# Funding

This research was supported by National Sciences and Engineering Research Council of Canada Discovery Grant RGPIN-2018-05255.

# Acknowledgements

The authors currently have no acknowledgements to make.

# Data Availability Statement

The code and datasets generated for this study may be found at <https://github.com/robwschlegel/MHWdetection>.

# Figure legends

Figure 1: The focus marine heatwaves (MHWs) shown in red for the three reference time series A) Western Australia (WA), B) Northwest Atlantic (NWA), and C) Mediterranean (Med). Other MHWs shown in salmon. Each panel is centred around the peak date of the focus MHW, which is highlighted by a dark green vertical segment. The beginning and end of each event are demarcated with light green vertical segments. The seasonal mean climatology for each time series is shown as a light blue line, while the threshold climatology is shown with a dark blue line. The observed temperatures are shown as a black line.

Figure 2: The effect of sub-optimal data on the MHWs detected in the final 10 years of time series. The columns show the results for each of the three sub-optimal tests: time series length (10 – 37 years), missing data (0 – 50%), and added decadal trends (0.00 – 0.30°C/dec). The rows show the three focus results from the MHW detection output: the change in the count of MHWs, the percentage of change in the sum of the MHW days, and the percentage change in the mean of the maximum intensities of the MHWs. Any samples that are significantly different (*p*-value <= 0.05) from the control are highlighted with red squares (there are none). The faint black lines in each panel show how the variable written on the y-axis changes as the data become more sub-optimal as shown on the x-axis as the sub-optimal tests were performed on 100 randomly selected pixels from the OISST dataset. The coloured lines show the effect of the sub-optimal tests on the three reference time series seen in Figure 1.

Figure 3: This figure shows results from the same set of tests visualised in Figure 2, but focusses only on the effects the sub-optimal tests had on the single largest event in the final 10 years of data in each time series. Significant differences (*p*-value <= 0.05) are highlighted in red (there are none). The top row of panels, “Count (n)”, shows the difference in the count of MHWs during what should be the duration of the focus event. A value of -1 means that no events were detected, and a value of 0 means that no additional MHWs were detected in addition to the single largest event. Theoretically this value should remain at 0, when it increases that means that the focus event is being chopped up into multiple smaller events. The bottom two rows of panels show percentage changes in the days of the focus MHW and its maximum intensity. A value 0f -100% means that no MHW was detected.

Figure 4: Global map showing changes in MHW detection as the time series at each pixel is shortened from 30 to 10 years. The left column shows the effect of time series length on all MHWs detected in the final 10 years of data, while the right column shows the effect on only the largest MHW. Panels A & D show the change in count of MHWs as the time series are shortened, panels B & E show the change in the duration (days) of the detected MHW(s), and panels C & F show the change in the maximum intensities (°C). The labels on the colour bars at the bottom of each panel show what the values are at the 5th, 10th, 50th, 90th, and 95th quantiles of the data shown in the coloured pixels. Any values smaller/larger than the 5th/95th quantile were rounded to prevent the very long tails from interfering with the visualisation of the results.

Figure 5: The effect of linear interpolation on the MHW results. These panels show the same information as Figure 2BEH after the missing data were filled via linear interpolation and the MHW detection algorithm was re-run. RWS: I’ll fix the text labels on this figure and add the effect on the focus events, too.

Figure 6: The count of consecutive missing days of data. RWS: I still need to create this figure.

Appendix 1: The effect of different 30 year climatology base periods on MHW results. RWS: I haven’t created this yet as I’m not certain it is necessary.

Appendix 2: The effect of the sub-optimal tests on the seasonal mean and threshold climatologies. These panels show the same information as Figure 3 except that the variables shown are the thresholds from the MHW detection output.

Appendix 3A: The same global map as Figure 4, but with rates of change shown due to increasing missing data from 0 – 50%.

Appendix 3B: The same global map as Figure 4, but with rates of change shown due to increasing decadal trends from 0.00 – 0.30°C/dec.

Appendix 4: The effect of changing window widths on MHW output as a fix for time series length. RWS: I have these results but haven’t made this figure yet as I don’t think it is necessary.

# References

Banzon, V., Smith, T. M., Chin, T. M., Liu, C., and Hankins, W. (2016). A long-term record of blended satellite and in situ sea-surface temperature for climate monitoring, modeling and environmental studies. *Earth System Science Data* 8, 165–176. doi:[10.5194/essd-8-165-2016](https://doi.org/10.5194/essd-8-165-2016).

Baumgartner, T. (1992). Reconstruction of the history of the pacific sardine and northern anchovy populations over the past two millenia from sediments of the santa barbara basin, california. *CalCOFI Rep* 33, 24–40.

Harrison, B., Jupp, D., Lewis, M., Forster, B., Mueller, N., Smith, C., et al. (2019). *Earth observation: Data, processing and applications*. Australia; New Zealand CRC for Spatial Information.

Hobday, A. J., Alexander, L. V., Perkins, S. E., Smale, D. A., Straub, S. C., Oliver, E. C., et al. (2016). A hierarchical approach to defining marine heatwaves. *Progress in Oceanography* 141, 227–238.

Oliver, E. C., Donat, M. G., Burrows, M. T., Moore, P. J., Smale, D. A., Alexander, L. V., et al. (2018). Longer and more frequent marine heatwaves over the past century. *Nature communications* 9, 1324.

Organization, W. M. (2011). *Guide to climatological practices*. World Meteorological Organization (WMO).

Organization, W. M. (2017). *WMO guidelines on the calculation of climate normals*. World Meteorological Organization (WMO).

Pachauri, R. K., Meyer, L., Ypersele, J. van, Brinkman, S., Kesteren, L. van, Leprince-Ringuet, N., et al. (2014). Climate Change 2014 Synthesis Report.

Philander, S. G. H. (1983). El nino southern oscillation phenomena. *Nature* 302, 295.

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., and Schlax, M. G. (2007). Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate* 20, 5473–5496.

Salinger, J., Hobday, A., Matear, R., O’Kane, T., Risbey, J., Dunstan, P., et al. (2016). “Chapter one - decadal-scale forecasting of climate drivers for marine applications,” in Advances in marine biology., ed. B. E. Curry (Academic Press), 1–68. doi:[https://doi.org/10.1016/bs.amb.2016.04.002](https://doi.org/https://doi.org/10.1016/bs.amb.2016.04.002).

Schlegel, R. W., and Smit, A. J. (2018). HeatwaveR: A central algorithm for the detection of heatwaves and cold-spells. *The Journal of Open Source Software* 3, 821.

Siegel, S., and NJ, C. (1988). *Non-parametric statistics for the behavioural sciences*. New York: McGraw-Hill.

Smit, A. J., Roberts, M., Anderson, R. J., Dufois, F., Dudley, S. F. J., Bornman, T. G., et al. (2013). A coastal seawater temperature dataset for biogeographical studies: large biases between in situ and remotely-sensed data sets around the coast of South Africa. *PLOS ONE* 8. doi:[10.1371/journal.pone.0081944](https://doi.org/10.1371/journal.pone.0081944).

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond “p<0.05”. *The American Statistician* 73, 1–19. doi:[10.1080/00031305.2019.1583913](https://doi.org/10.1080/00031305.2019.1583913).

Zhao, Z., and Marin, M. (2019). A matlab toolbox to detect and analyze marine heatwaves. *The Journal of Open Source Software* 4, 1124.