

Toward statistically optimal temperature extreme detection

Zijie Zhao

`zijie.zhao@utas.edu.au`

University of California Irvine <https://orcid.org/0000-0003-3403-878X>

Alistair J. Hobday

CSIRO Environment

Neil J. Holbrook

University of Tasmania

Alex Sen Gupta

UNSW Sydney

Eric C. J. Oliver

Dalhousie University

Lukas Brunner

Universität Hamburg

Michael T. Burrows

Scottish Association For Marine Science

Karen Filbee-Dexter

University of Western Australia

Neil Malan

UNSW Sydney

Shinae Montie

University of Western Australia

Pippa J. Moore

Newcastle University

Dan A. Smale

Marine Biological Association of the United Kingdom

Kathryn E. Smith

Marine Biological Association of the United Kingdom

Mads S. Thomsen

`mads.thomsen@canterbury.ac.nz`

University of Canterbury

Thomas Wernberg

University of Western Australia

Research Article

Keywords:

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11328032/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Toward statistically optimal temperature extreme detection

Zijie Zhao^{a,b,c,*}, Alistair J. Hobday^d, Neil J. Holbrook^{a,c,e}, Alex Sen Gupta^f, Eric C. J. Oliver^g,
Lukas Brunner^h, Michael T. Burrowsⁱ, Karen Filbee-Dexter^{j,k}, Neil Malan^f, Shinae Montie^j,
Pippa J. Moore^l, Dan A. Smale^m, Kathryn E. Smith^m, Mads S. Thomsen^{j,n,o}, Thomas
Wernberg^{j,k}

^a Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania, Australia.

^b Department of Earth System Science, University of California, Irvine, CA, USA.

^c ARC Centre of Excellence for Climate Extremes, University of Tasmania, Hobart, Tasmania, Australia

^d CSIRO Environment, Hobart, Tasmania, Australia.

^e ARC Centre of Excellence for the Weather of the 21st Century, University of Tasmania, Hobart, Tasmania, Australia.

^f Climate Change Research Centre, Centre for Marine Science & Innovation and ARC Australian Centre for Excellence in Antarctic Science, University of New South Wales, Sydney, Australia.

^g Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada.

^h Research Unit Sustainability and Climate Risk, Earth and Society Research Hub (ESRAH), University of Hamburg, Hamburg, Germany.

ⁱ Scottish Association for Marine Science, Oban, United Kingdom.

^j UWA Oceans Institute & School of Biological Sciences, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia.

^k Norwegian Institute of Marine Research, Nye Flødevigveien 20, 4817, His, Norway.

^l Dove Marine Laboratory, School of Natural and Environmental Sciences, Newcastle University, Newcastle-Upon-Tyne, UK.

^m The Marine Biological Association of the UK, Citadel Hill, Plymouth, UK.

ⁿ Marine Ecology Research Group, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.

^o Department of Ecoscience, Aarhus University, Roskilde, Denmark

Corresponding author: Zijie Zhao, zijiz26@uci.edu

ABSTRACT

Temperature extremes are commonly detected relative to climatological baselines estimated from short observational records, but the statistical reliability of these definitions remains poorly constrained. We develop an observation-constrained framework for objectively evaluating and optimizing temperature-extreme detection from finite records. Using marine heatwaves as a case study, we first compare uncorrected moving-window methods with anomaly-based corrected methods and show that the conventional approach introduces systematic biases in the mean occurrence and seasonal amplitude of detected events. The corrected formulation substantially reduces these biases. We then generate large ensembles of synthetic global sea surface temperature time series constrained by observations. These synthetic records reproduce key observed statistical properties, including seasonal structure, temporal persistence, and large-scale variability, and provide a long reference state against which detection choices can be tested. Defining optimality as the minimization of errors in marine-heatwave severity relative to this reference state, we show that optimal moving-window lengths vary with record length and regional dynamics. For satellite-era sea surface temperature records, a 31-day window provides a robust global choice. Applying optimized and conventional definitions to observations demonstrates that methodological choices can substantially alter inferred marine-heatwave metrics, with discrepancies increasing rapidly for shorter records and becoming largest for baselines shorter than about 15 years. The framework provides a reproducible basis for selecting detection methods and baseline parameters when temperature extremes must be defined from limited observations.

SIGNIFICANCE STATEMENT

Temperature extremes are often used to describe climate risk, but their definition can depend on methodological choices made when estimating climatological baselines from short records. Using marine heatwaves as an example, we show that these choices can bias the number and seasonal distribution of detected events. We then create long synthetic sea surface temperature records constrained by observations to test which definitions best recover known extreme-event statistics. This approach identifies more reliable detection choices, including a 31-day window for satellite-era sea surface temperature records, and shows that errors become especially large when baselines are shorter than about 15 years. The method helps make temperature-extreme detection more objective and reproducible.

1. Introduction

In many recent studies, extreme temperature events are identified using percentile-based thresholds relative to a climatological baseline (Hobday et al. 2016; Li et al. 2024; Liu et al. 2023b; Perkins; Alexander 2013; Schlegel et al. 2021; Thorat et al. 2026; Woolway et al. 2021; Zhang et al. 2011; Zscheischler et al. 2020). While conceptually straightforward, this approach implicitly assumes that the baseline can be estimated accurately from observational records that are often short, noisy, and incomplete (Schlegel et al. 2019). In practice, this assumption is rarely satisfied, raising a fundamental question when detecting extreme events from thresholds: how should extremes be defined when the underlying climate statistics are only partially observed?

To mitigate sampling noise, moving-window approaches are widely used to estimate seasonally varying means and percentile-based thresholds. These methods increase the effective sample size by pooling data across adjacent calendar days and have become standard practice in the detection of a wide range of extremes, including atmospheric heatwaves (Freychet et al. 2021; Perkins; Alexander 2013; Russo et al. 2015; Schlegel et al. 2021; Vogel et al. 2020), marine heatwaves (MHWs) (Hobday et al. 2016; Pearce et al. 2011), cold spells (Schlegel et al. 2021), and other oceanic and hydrological events (Liu et al. 2023a; Woolway et al. 2021). Despite their ubiquity, the choice of window length is often arbitrary and subjective in related studies, lacking objective criteria for determining an optimal size. Consequently, the statistical properties of moving windows, and the implications of window-length selection, remain insufficiently characterized. As a result, critical elements in the definition of extreme events continue to rely on heuristic conventions rather than quantitative criteria for optimality.

MHWs (Hobday et al. 2016; Pearce et al. 2011) have emerged as one of the most prominent climate extremes of recent decades in both public discourse and academic research due to their devastating impacts on marine biodiversity and ecosystem services (Smale et al. 2019; Smith et al. 2021; Smith et al. 2023; Smith et al. 2024; Wernberg et al. 2025), which provide a particularly clear example of the extreme-definition problem. Identified as discrete and prolonged warm ocean temperature extremes, MHWs are typically identified using a seasonally varying 90th-percentile threshold derived from records spanning only a few decades, with moving windows commonly employed to stabilize the baseline. However, Brunner and Voigt (Brunner; Voigt 2024) demonstrated, using atmospheric heatwaves as a case study, that moving-window-based definitions can introduce systematic biases in both the

frequency and seasonal distribution of detected extremes, even when thresholds are designed to account for seasonality. They proposed a corrected approach (BV-Corrected; Fig. 1), in which the seasonal 90th-percentile baseline is derived from temperature anomalies, after removal of the seasonal mean, rather than from raw time series, thereby reducing such biases. Nevertheless, although the BV-Corrected approach substantially reduces biases arising from the use of moving windows, it does not establish a general criterion for selecting the optimal window size in heatwave detection, leaving a key methodological gap unresolved.

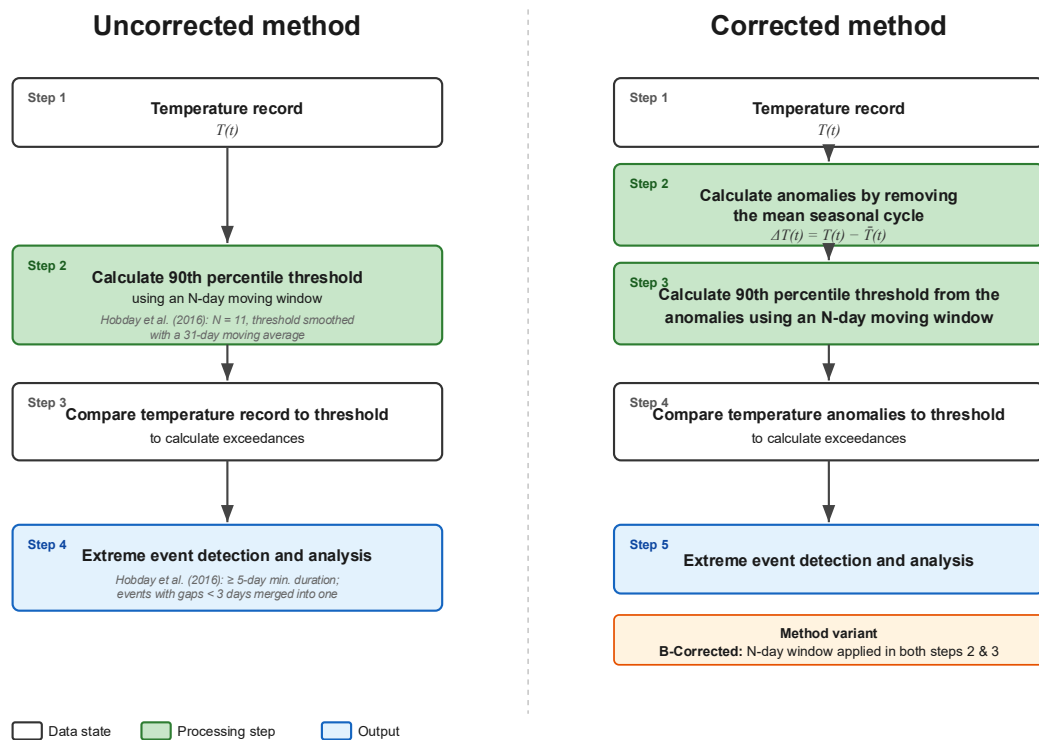


Fig. 1. Protocols for uncorrected and corrected methods to detect extreme temperature events. The uncorrected method first applies a moving window to the temperature record to establish a baseline and then identifies and quantifies extreme events based on that baseline. In contrast, the corrected method (BV-Corrected) begins by removing the climatological mean (seasonal cycle) from the temperature record, calculates percentile thresholds from the resulting anomalies, and subsequently detects extremes. A moving window is then used when computing percentile thresholds from the anomaly time series. When the same moving window is applied to both steps 2 and 3, the method corresponds to the B-Corrected variant.

A major obstacle is the absence of a reliable reference against which extreme-event definitions can be evaluated. Global satellite observational records (from 1981) are still too

short to robustly estimate “true” climatological baselines, while large ensembles from climate models are computationally expensive and subject to model-dependent biases in simulating extremes. Without a reliable and unbiased long term reference dataset, it is not possible to objectively assess detection errors or to determine optimal methodological choices.

Here we address this limitation by developing an observation-constrained framework for statistically consistent and objectively optimized extreme-event detection. Using MHWs as a case study, we treat detection as a joint problem involving both the choice of detection method and the parameterization of climatological baselines, specifically the moving-window length here. We first quantify biases associated with uncorrected and corrected MHW detection methods and show that a fully corrected formulation substantially improves statistical consistency, thereby identifying a more robust detection method. We then construct large ensembles of synthetic global sea surface temperature (SST) time series constrained by observations and designed to reproduce realistic statistical properties, including seasonal structure, temporal persistence, and large-scale climate variability. These synthetic records provide a statistically robust approximation of a “true” climate state, enabling direct evaluation of detection methods under controlled conditions. Within this framework, we define parameter optimality as the minimization of errors in MHW severity relative to the synthetic reference and use this criterion to determine the optimal moving-window length as a function of record length and regional dynamics. Finally, we compare MHW metrics estimated using the conventional and optimally corrected methods, showing that substantial discrepancies arise when the available records are limited.

2. Data and Method

a. Sea surface temperature observations

Global daily SST during 1982-2022 were extracted from the NOAA OI SST V2 (Reynolds et al. 2007), which has been widely used as a benchmark observational dataset for global MHW analyses. The dataset used in this study was regridded to different spatial resolutions according to the requirements of each analysis. For the comparison of MHW detection methods in section 3, the data were regridded from the original 0.25° resolution to 1°. For the generation of large ensembles of synthetic SST time series and the corresponding MHW detection analyses in sections 4 and 5, the data were regridded to 2.5° because of computational constraints. For the

comparison of MHW metrics derived from the conventional and optimally corrected methods in section 6, the original 0.25° grid was retained.

b. Methods to detect marine heatwaves

The uncorrected or conventional method used here for detecting MHWs identifies temperature events exceeding a seasonally varying daily 90th percentile threshold. This threshold was calculated using a moving window of length N days. For instance, applying an 11-day window across a 30-year temperature record means the 90th percentile for each Julian day is based on $11 \times 30 = 330$ temperature observations. In effect, the uncorrected approach described here mirrors the Hobday et al. definition but omits two components: the requirement that events persist for at least five consecutive days with gaps less than two days, and the application of a 31-day moving window to further smooth the percentile-based threshold. Here we consider a fixed baseline rather than a moving baseline (Amaya et al. 2023) to align with the analytic process by Brunner and Voigt (Brunner; Voigt 2024).

In contrast to the uncorrected approach, the BV-corrected method does not directly calculate the 90th percentile threshold from raw temperature records. Instead, it first removes the climatological seasonal cycle, computed as the mean temperature for each calendar day, from the raw temperature series. Temperature anomalies derived from this procedure were used to estimate the seasonal 90th percentile baseline, with an optional moving window applied. MHWs are thus identified when temperature anomalies exceed the day-of-year varying percentile threshold.

Depending on how moving windows are incorporated into the daily climatological mean calculation, a variant of the BV-corrected method can be distinguished, the B-Corrected method, in which moving windows are applied to both the seasonal mean and 90th percentile estimation (Fig.1).

To quantify biases introduced by MHW detection methods, we focus on two key metrics: the mean and seasonal amplitude of MHW proportions. Under idealized conditions, MHWs are assumed to occur with a theoretical frequency of 10% and exhibit no seasonal cycle (i.e., seasonal amplitude = 0). In reality, however, MHWs may still display seasonal modulation because some seasonal variance can remain even after removing the seasonally varying mean. Even in such condition, the deviation of the observed MHW seasonal amplitude from zero can still serve as a proxy for assessing the influence of the moving-window approach.

The mean MHW proportion was calculated as the percentage of MHW occurrences over the full analysis period (1982–2022). Similarly, monthly proportion was computed as the percentage of MHW occurrences within each calendar month. A harmonic regression was then applied to the resulting monthly time series using the equation:

$$M = a + b \sin\left(\frac{2\pi}{T}t + \varphi\right),$$

where M is the monthly MHW proportion, a is the mean offset, b is the seasonal amplitude, $T=12$ (months), and φ is the phase term.

Biases in mean and seasonal amplitude of MHW proportion are estimated as the deviations from their respective theoretical values, i.e., the difference between the observed mean frequency and 10%, and between the observed seasonal amplitude and 0.

c. Simulate sea surface temperature in a synthetic world

We aimed to develop a method for producing synthetic SST time series that preserve key statistical properties of the observed record, including the seasonal mean and variance, short- and long-term autocorrelations, and low-frequency fingerprints associated with large-scale climate modes such as the El Niño–Southern Oscillation (ENSO). To achieve this, synthetic SST were generated by separating observed variability into a large-scale, dynamically driven component and a local, stochastic component. The large-scale variability was simulated using a linear inverse model (LIM) (Penland; Sardeshmukh 1995), while the local variability was standardized and represented using autoregressive processes. The two components were then recombined and rescaled to recover the observed seasonal structure, producing long synthetic time series that preserve key statistical and dynamical characteristics of the observations. Details of the process are described as follows.

1) PREPARATION FOR SST SYNTHESIS

Before generating synthetic SST, we first estimated several metrics required during the synthesis process, including the seasonal mean, the seasonal cycle of variance, and daily SST anomalies (SSTA). Let $t = 1, 2, \dots, T$ denote the daily time index and the corresponding calendar day.

The seasonal mean SST was represented by fitting the first three harmonics:

$$\mu_{\text{season}}(d(t)) = \beta_0 + \sum_{k=1}^3 \left[a_k \cos\left(\frac{2\pi k d(t)}{P}\right) + b_k \sin\left(\frac{2\pi k d(t)}{P}\right) \right], P = 365.2425. \quad 1$$

A linear trend was modelled as

$$\tau(t) = \alpha_0 + \alpha_1(t - \bar{t}). \quad 2$$

Daily SSTA were then defined as

$$\text{SSTA}(t) = \text{SST}(t) - \mu_{\text{season}}(d(t)) - \tau(t). \quad 3$$

SSTA variance can show important seasonal differences, for example related to changes in mixed layer depth. We estimated the seasonal cycle of variance by computing the daily SSTA variance and fitting its logarithm with the first three harmonics

$$\log \sigma^2(d(t)) = \gamma_0 + \sum_{k=1}^3 \left[c_k \cos\left(\frac{2\pi k d(t)}{P}\right) + d_k \sin\left(\frac{2\pi k d(t)}{P}\right) \right]. \quad 4$$

Standardized anomalies were obtained as

$$z(t) = \frac{\text{SSTA}(t)}{\sigma(d(t))}. \quad 5$$

Here, we estimate the seasonal mean and variance using three harmonics, a convenient and widely used choice for representing climatology. A potentially more optimal approach would be to fit harmonics stepwise until the addition of further terms is no longer statistically significant using the method described in Appendix A. Although this procedure may select a larger number of harmonics, it produces negligible changes in the fitted seasonal cycle (Fig. A1). We therefore retain three harmonics to ensure a simple, consistent criterion across datasets and future projections, rather than adopting data-dependent method complexity.

2) SYNTHESIS OF THE REMOTE LOW-FREQUENCY COMPONENT OF SSTA

To synthesize the remote component, we first converted daily SSTA into monthly anomalies via monthly averaging. Large-scale, linear SSTA variability was then simulated using a LIM:

$$\frac{dX}{dt} = LX + \xi, \quad 6$$

where $X(t)$ is the climate state represented by empirical orthogonal function (EOF) modes, L is the linear dynamical operator, and ξ is spatially structured white noise. Conceptually, the LIM behaves as a spatially aware autoregressive model, capturing both the temporal autocorrelation and spatial covariance structure of large-scale climate variability.

We retained the leading 20 principal components (PCs), which together explained 46.1% of total variance, and estimated L from the one-month lag covariance. After fitting the LIM, we generated simulated remote SST components by integrating the model forward with random initial conditions. These simulations produced an ensemble of monthly remote SST anomalies, denoted $SSTA_{remote}$, which represent the projection of SST anomalies onto the leading 20 EOF modes ($SSTA_{EOF}$). We used only the leading 20 PCs because including a larger number of PCs to reconstruct the remote component of SST anomalies can bias the synthetic SST toward an overestimation of the observed autocorrelation. This effect is illustrated by comparing a series of synthetic SST anomaly whose remote components are reconstructed from 20 versus 100 PCs (Appendix B; Fig. B1). We hypothesize that the bias arises from the monthly data used to construct the LIM. Monthly data were required because daily data are often too noisy to preserve the approximately linear dynamics assumed by the LIM. However, relative to daily data, monthly averages naturally retain stronger correlations at longer lags because much of the high-frequency daily variability is smoothed out. Consequently, reconstructing the remote component with many PCs more fully captures this monthly-scale memory and can transfer excessive persistence into the daily synthetic SST anomalies. We therefore used 20 PCs as a conservative choice that balances the need to capture the dominant remote forcing patterns against the risk of introducing excessive monthly memory into daily-scale variability

3) SYNTHESIS OF THE LOCAL COMPONENT OF SSTA

To isolate locally generated variability, we removed the EOF-based remote signal from the full SSTA:

$$SSTA_{res} = SSTA - SSTA_{EOF}. \quad 7$$

We then standardized this residual component by its own seasonal variance:

$$z_{res}(t) = \frac{SSTA_{res}(t)}{\sigma_{res}(d(t))}, \quad 8$$

yielding an approximately stationary series with unit variance.

Short-timescale persistence was represented by fitting an autoregressive-moving average model, ARMA(1,1), to $z(t)$:

$$S_{\text{short}}(t) = \phi S_{\text{short}}(t-1) + \varepsilon_t + \theta \varepsilon_{t-1}, \varepsilon_t \sim N(0, \sigma_\varepsilon^2). \quad 9$$

This captures day-to-day autocorrelation curvature but not the longer persistence associated with ocean thermal memory. To represent the slow (long-term) component, we construct a simple first order autoregressive model, i.e., AR(1) process:

$$S_{\text{long}}(t) = \rho S_{\text{long}}(t-1) + \eta_t, \eta_t \sim N(0,1), \quad 10$$

which provides a red-noise background for long-timescale memory.

Directly fitting Eq. (10) to $z(t)$ captures autocorrelation only at day-to-day timescales and may fail to represent memory effects operating over longer timescales. We therefore estimated the memory coefficient ρ by constructing a variance-preserving mixture of short- and long-timescale components, rather than by direct fitting:

$$S_{\text{mix}}(t) = \sqrt{1-f} S_{\text{short}}(t) + \sqrt{f} S_{\text{long}}(t), f \in [0,1]. \quad 11$$

We sweep over (ρ, f) and selected optimal values (ρ^*, f^*) that minimize the mismatch between the autocorrelation functions (ACFs) of $z_{\text{res}}(t)$ and $S_{\text{mix}}(t)$ at long lags:

$$(\rho^*, f^*) = \arg \min_{\rho, f} \sum_{L=L_1}^{L_2} [\text{ACF}_S(L) - \text{ACF}_z(L)]^2. \quad 12$$

We used $L_1 = 10$ and $L_2 = 60$ days because autocorrelations at shorter lags are already captured by the ARMA component, and ρ was evaluated from 0.03 to 0.99. The resulting optimal mixture defines the normalized local component:

$$z_{\text{local}} = S_{\text{mix}}(\rho^*, f^*). \quad 13$$

4) COMBINING REMOTE AND LOCAL COMPONENTS

To ensure compatibility, we standardized the remote component:

$$z_{\text{remote}}(t) = \frac{\text{SSTA}_{\text{remote}}(t)}{\sigma_{\text{remote}}(d(t))}. \quad 14$$

The normalized synthetic SSTA were obtained by forming a weighted mixture of the local and remote components, similar to Eq.12:

$$z_{\text{mix}}(t) = \sqrt{1-h} z_{\text{remote}}(t) + \sqrt{h} z_{\text{local}}(t), h \in [0,1]. \quad 15$$

The optimal weight h^* was chosen to minimize the mismatch between the ACFs of $\sigma(d(t))z_{\text{mix}}(t)$ and the observed standardized anomalies:

$$h^* = \arg \min_h \sum_{L=1}^{60} [\text{ACF}_{\text{mix}}(L) - \text{ACF}_{\text{obs}}(L)]^2. \quad 16$$

The normalized synthetic SSTA are thus

$$z_{\text{syn}} = z_{\text{mix}}(h^*). \quad 17$$

Finally, we rescaled by the seasonal standard deviation to recover the seasonal modulation of variability:

$$\text{SSTA}_{\text{syn}}(t) = \sigma(d(t)) z_{\text{syn}}(t), \quad 18$$

and reconstructed the full synthetic SST by adding back the seasonal cycle:

$$\text{SST}_{\text{syn}}(t) = \mu_{\text{season}}(d(t)) + \text{SSTA}_{\text{syn}}(t). \quad 19$$

Using the method described in the previous section, we synthesized an ensemble of 1,000 global SST records, each spanning 5,000 years, constrained by observed global SST variability during 1982–2022. Given the high computational demands, this process is conducted on regridded 2.5° grid cells.

d. Determination of optimal window sizes used in MHW detection

Because MHWs are characterized by both binary occurrence (presence or absence) and intensity (SSTA), relying solely on occurrence to determine the optimal window size may introduce bias. We therefore used the MHW severity index (SI) (Sen Gupta et al. 2020) to identify the optimal window size, which is calculated as:

$$SI(t, w, y) = \frac{\text{SSTA}(t, w, y)}{a90(d(t), w, y)},$$

where $SSTA(t, w, y)$ denotes the SSTA relative to the seasonal mean estimated using a window of size w and the first y years of synthetic SST, and $a_{90}(d(t), w, y)$ represents the seasonal 90th percentile of $SSTA(t, w, y)$ computed for each calendar day $d(t)$. The optimal window size, w_{opt} , for a given y is defined as the window length that minimizes the root mean square error (RMSE) between the resulting SI and SI_{true} , where SI_{true} is calculated using baselines derived from the full 5,000-year synthetic SST record without windowing.

Besides the abovementioned manner, the SI can also be calculated regarding baselines constructed using the uncorrected approach, where seasonal mean and 90th percentile are derived from raw SST time series independently. In such case, the SI is written as:

$$SI(t, w, y) = \frac{SSTA(t, w, y)}{p90(d(t), w, y) - seas(d(t), w, y)},$$

where $p90(d(t), w, y)$ and $seas(d(t), w, y)$ indicates the seasonal 90th percentile calculated using a window of size w and the first y years of synthetic SST directly.

A further advantage of using the SI to quantify MHW occurrence is its compatibility with our experimental design. We excluded long-term trends from the synthetic time series because the procedure for determining the optimal window length requires a fixed number of years to define the climatology. When a trend is present, the estimated climatology—and hence any binary measure of MHW occurrence based on it—depends on the location of the selected interval within the full record. This dependence can inflate the apparent occurrence of MHWs in later years, violating the stationarity assumption underlying the experimental design. SI circumvents this problem because it is a continuous metric defined throughout the record: a trend affects its magnitude but not its availability across different phases of the time series. Consequently, an optimal window derived from trend-free synthetic records can remain applicable under realistic, nonstationary conditions. Consistent with this expectation, a sensitivity analysis (Appendix C) in which linear trends of varying magnitudes were added to the synthetic SST time series showed that the identified optimal window remained stable (Fig. C1). This framework therefore provides an internally consistent basis for determining optimal window lengths that can be transferred to analyses of real-world records.

3. Validity of MHWs detection using the three methods

While Brunner and Voigt (2024) proposed removing the seasonal mean prior to threshold calculation to mitigate bias, they did not specify if the seasonal mean should be determined

with a moving window. For the purpose of constructing an optimal “true” baseline, incorporating a moving window into the seasonal mean calculation has the potential to stabilize the baseline and thereby enhance the detection of MHWs. By incorporating a moving window into this step, a variant of the corrected method can be defined—applying it to both seasonal mean and percentile estimation (B-Corrected)—yielding three methods in total: uncorrected, BV-, and B-Corrected (Fig. 1). The ability of these corrected variants to generate accurate MHW statistics in identifying MHWs therefore requires further assessment.

We first assessed MHW detection across global ice-free regions using the three methodological variants (uncorrected, BV-corrected, and B-corrected), each employing an 11-day moving window, in construction of 90th percentile thresholds for BV-corrected or both percentile threshold and climatological mean for B-corrected (Fig. 1). Here we adopt a simplified definition of MHWs that does not impose a minimum duration threshold or apply post-processing to calculated baselines. Under this definition, when the analysis period

coincides with the climatological baseline period, a grid cell will, by construction, experience MHW conditions for approximately 10% of the total time.

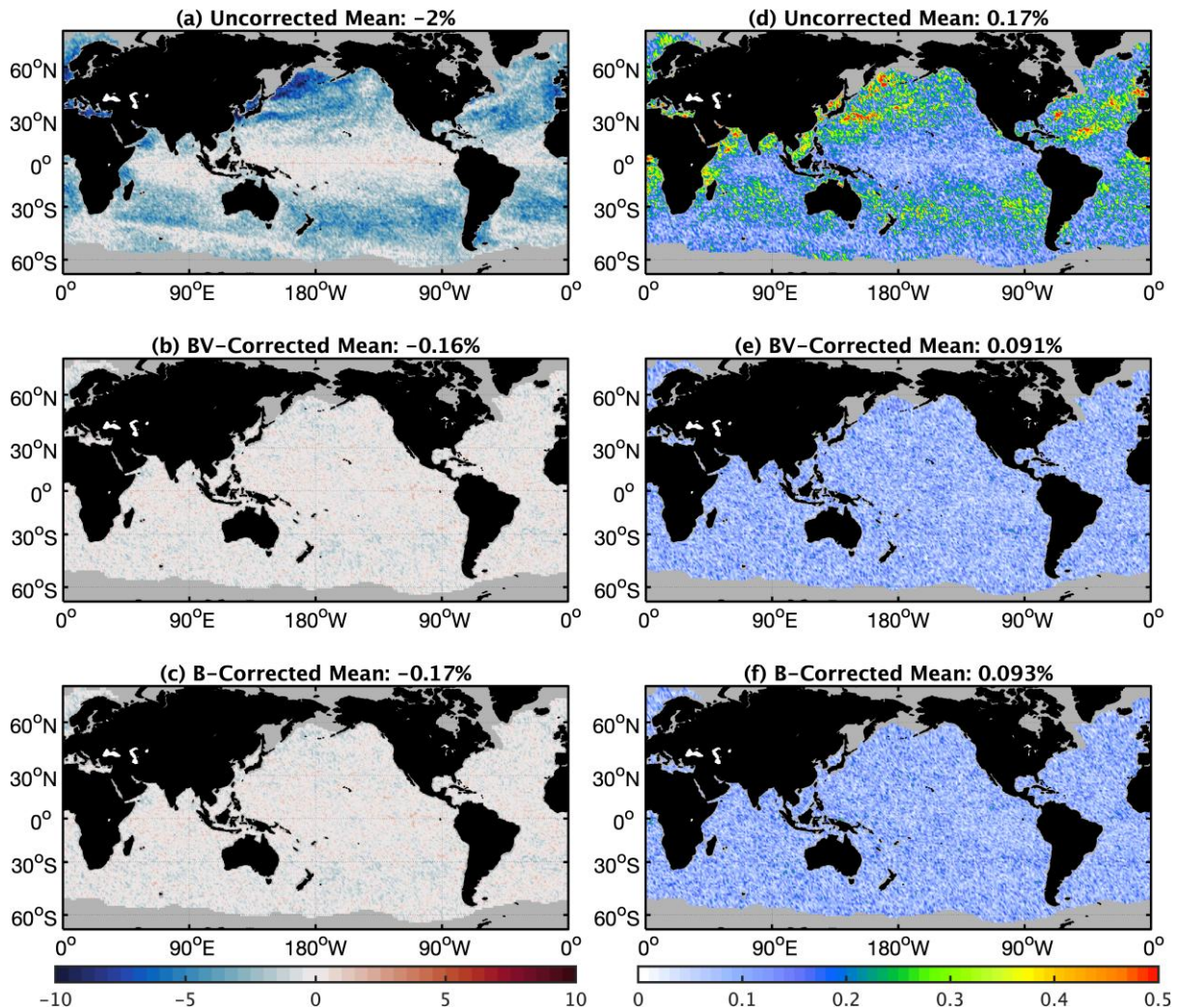


Fig. 2. Spatial biases and seasonal amplitudes of MHW proportion using three methodological variants across global ice-free regions. (a–c) Spatial biases in MHW proportion relative to the theoretical 10% expectation, and (d–f) seasonal amplitude of MHW frequency, derived from four methodological variants with an 11-day moving window: (a, d) Uncorrected, (b, e) BV-Corrected (moving window applied during percentile threshold calculation), and (c, f) B-Corrected (moving window applied during both climatology and percentile threshold calculation).

The uncorrected method exhibited widespread spatial biases relative to the theoretical proportion of MHW occurrences (10%), with a global mean underestimation of 2.0% and localized deviations reaching up to 20% in regions such as the Bering Sea, Kuroshio Extension, and North Atlantic (Fig. 2a). These biases showed pronounced spatial variability, tending to be

larger at high latitudes and smaller at low latitudes. This pattern resembles the global SST seasonal amplitude distribution (Levitus 1987), suggesting the uncorrected method produces larger biases in regions with strong seasonal cycles, likely because more seasonal information is transferred to resultant MHW outputs through moving windows. The two correction methods substantially reduced these biases, with global mean underestimations an order of magnitude smaller and a more spatially uniform bias distribution (from 2% to <0.2%, Fig. 2a,c,e).

Since MHWs here are defined relative to a seasonally varying climatological baseline (Smith et al. 2025), the annual cycle of their MHW proportion is expected to exhibit minimal seasonality with only weak seasonal modulations from seasonally varying variance. However, both the findings of Brunner and Voigt (2024) and our results (Fig. 2d) indicate that pronounced seasonality persists in the annual MHW occurrence cycle when using the uncorrected method. In regions where MHW frequency biases are prominent, the seasonal amplitude of MHW proportion can exceed 0.50% (i.e., more than a 1% difference between the annual maximum and minimum, corresponding to one-tenth of the uncertainty relative to the theoretical 10% frequency; see Section 2 for details), with a global mean of 0.17% (Fig. 2e). Applying the BV- and B-corrected methods reduces the global mean amplitude to 0.091% and 0.093%, respectively, while substantially suppressing spatial variability, indicating that the residual seasonality is reasonably mitigated.

We investigated the impact of varying moving window sizes on MHW detection by assessing global mean MHW frequency and its seasonal amplitude. MHWs were identified across the global ice-free oceans using window sizes ranging from 1 day (no window) to 31 days, for each of the three different methods (uncorrected, BV-corrected, B-corrected, see Fig. 1). We then evaluated the global mean and spatial standard deviation of MHW frequency and its seasonal amplitude. Under the Uncorrected method, the global mean MHW proportion decreased as window size increased, declining from 9.8% with an 11-day window to 8.8% with a 31-day window (Fig. 3a). Simultaneously, the mean seasonal amplitude increased from 0.17% to 0.68% (Fig. 3b), and the spatial standard deviation of both metrics increased with window size (Fig. 3c). In contrast, the BV- and B-Corrected methods yielded stable or declining global means and spatial variability for both MHW proportion and seasonal amplitude as window size increased (Fig. 3a-d). These results indicate that the corrected

methods yield less biased estimates of both the mean and seasonal proportion of MHWs, thereby providing a more robust framework for MHW detection

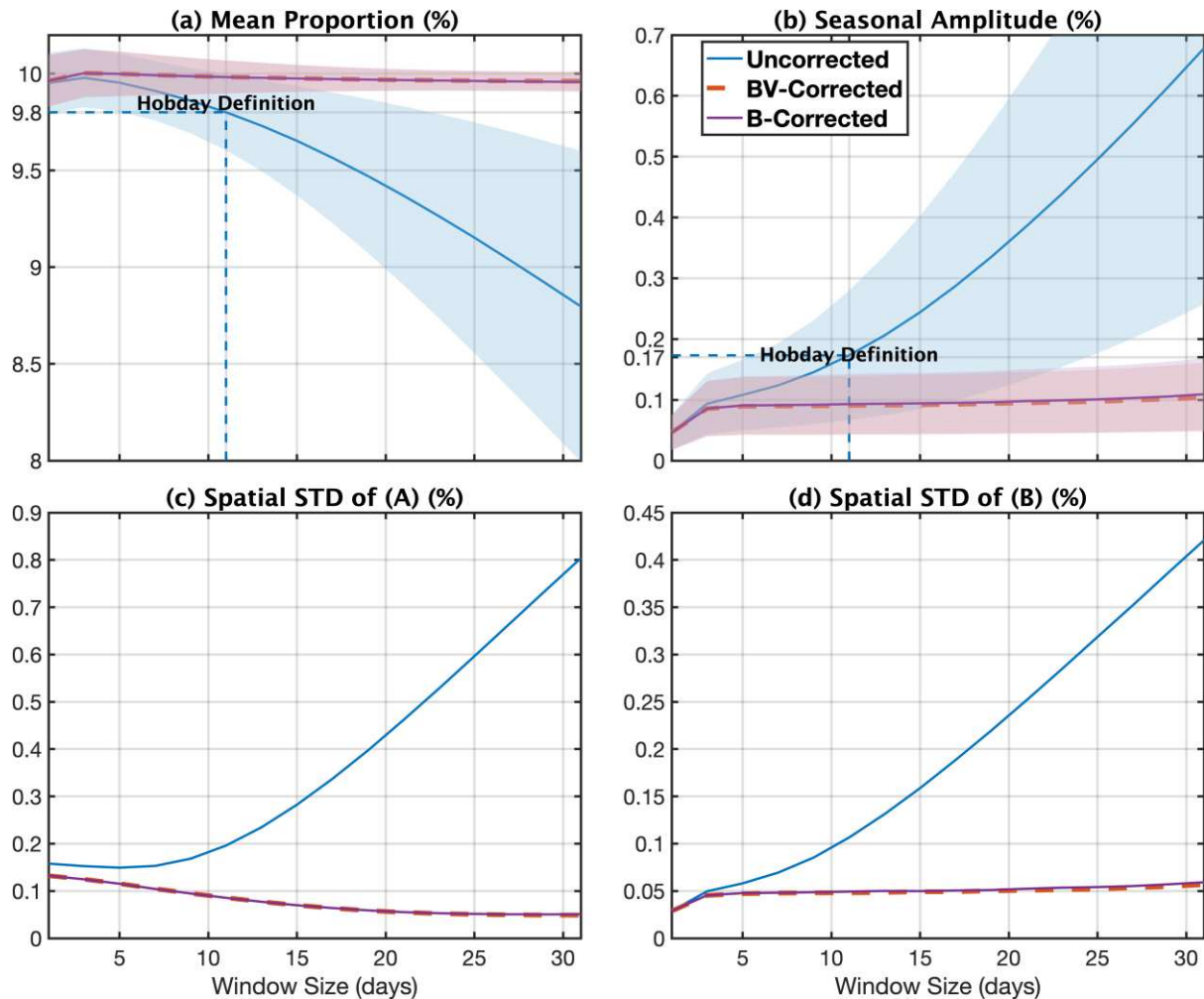


Fig.3. Sensitivity of MHW detection to moving window size. Global (a) mean proportion and (b) seasonal amplitude of MHWs, and (b,d) their spatial standard deviations, respectively, across moving window sizes from 1 (no moving window applied) to 31 days. Lines represent different methodological variants; shaded areas indicate ± 1 standard deviation. Results based on the window size specified by Hobday et al. (2016) are shown. The BV-corrected curve (orange, dashed) is distinguished from the B-corrected curve (purple, solid).

4. Identifying MHWs in a synthetic world

When detecting extremes from relatively short records—typically 30–50 years for global SST satellite observations—baselines are commonly estimated using moving windows to produce smoother climatologies, thereby avoiding noise being spuriously added to the derived

extremes. However, the choice of window length has been largely arbitrary in previous studies, and an objective criterion for its determination is lacking.

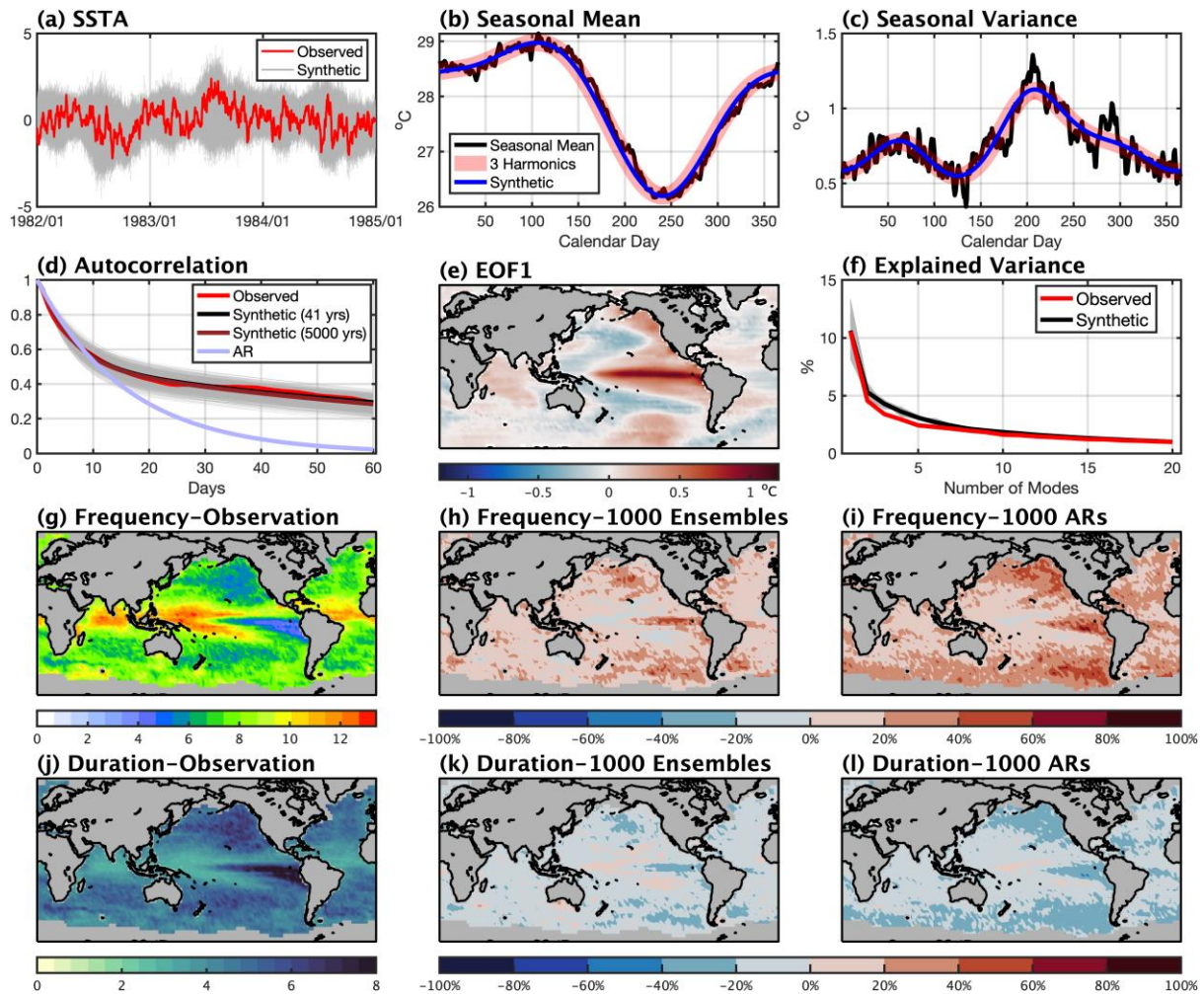


Fig.4. Synthetic sea surface temperature and marine heatwave simulations in an idealized framework. SST is simulated over 5,000 years using the method described in Methods. (a) Three-year observed SST at [106.25° E, 8.75° S] compared with synthetic SST time series from 1,000 independent simulations (each 5,000 y). (b) Seasonal-mean SST from 41 y of observations (black), its fit using the first three harmonics (red), and corresponding means from the 5,000-year simulations (blue). (c) As in b, but for seasonal variance. (d) Autocorrelation functions at lags of 1–60 d for observed and synthetic SST. The synthetic autocorrelation is stable whether estimated from a random 41-year subset or the full 5,000-year record, indicating robust statistical properties; autocorrelation from SST anomalies simulated using an AR(1) model forced by local autocorrelation and SSTA standard deviation is shown for reference. (e) Leading empirical orthogonal function (EOF1) patterns composited across the 1,000 simulations. (f) Variance explained by the first 20 EOF modes compared with

observations. (g) Observed annual MHW frequency (events year⁻¹) from 41 years of observations and percentage change (%) relative to synthetic estimates from 1,000 ensembles of 41-years (h) synthetic SST and (i) AR(1)-simulated SST. (j–l) As in G–I, but for MHW duration (days).

To evaluate the performance of moving-window approaches under limited baseline conditions, we generated 1,000 ensembles of synthetic SST time series, each spanning 5,000 non-leap years across the global ocean (Fig. 4). The synthetic SST reproduces key observed statistical properties, including the long-term mean (Fig. 4b), seasonally varying variance (Fig. 4c), and the autocorrelation structure at daily to subseasonal timescales (1–60 days; Fig. 4d; see methods) across diverse regions, including a coastal grid cell in the western Indian Ocean, the North Atlantic, the Tasman Sea, and the eastern Pacific (Fig. 4a–d; Fig. 5). Notable biases are only confined to the grid cell in the North Atlantic, where autocorrelation at 10–40-day lags is weaker than observed (Fig. 5d). Nevertheless, the synthetic time series more accurately captures SST internal memory than autoregressive models constrained by local autocorrelation and SST anomaly variance (Fig. 4d; Fig. 5d,h,l). The synthetic fields are constructed to preserve large-scale climate variability, as represented by the leading EOFs (Fig. 4e,f). In particular, the ENSO pattern is well captured in the leading EOF mode, with comparable

explained variance to observations, although variance is slightly underestimated in modes 2–5.

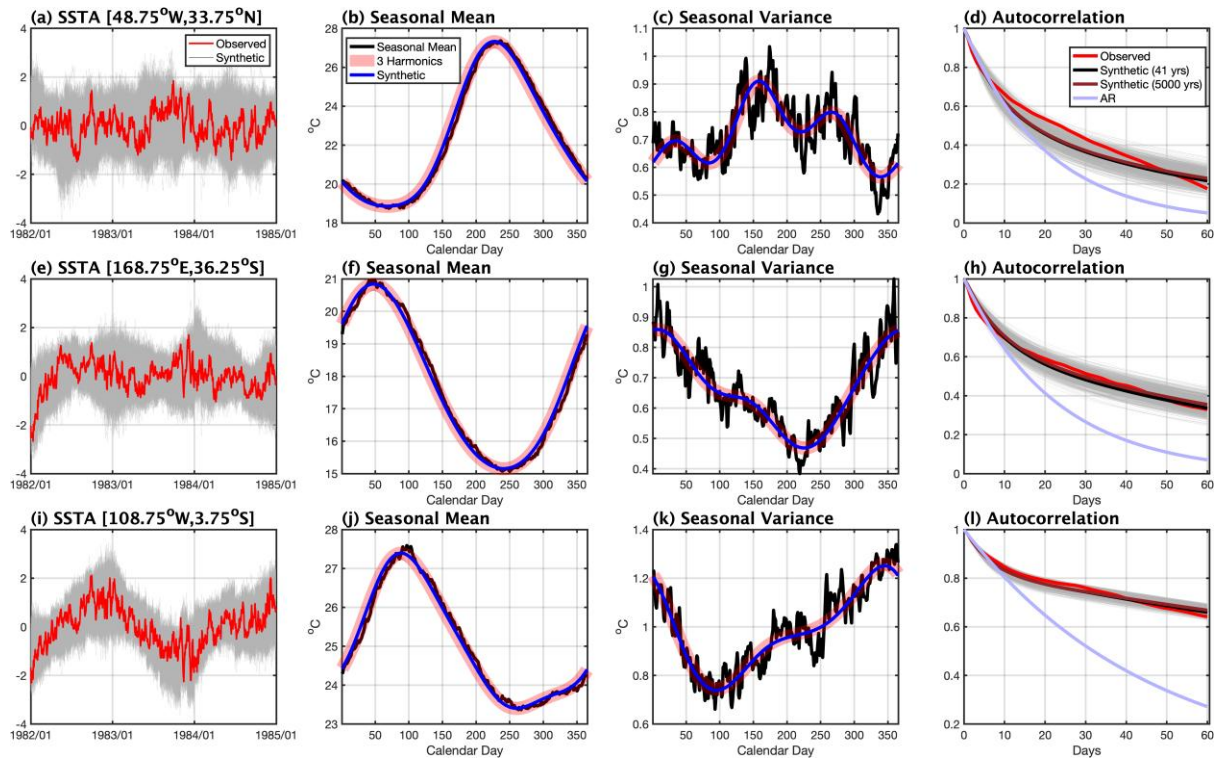


Fig. 5. Synthetic sea surface temperature across different regions. Similar to Fig. 4a–d, but based on synthetic SST ensembles from selected grid cells in (a–d) the North Atlantic, (e–h) the Tasman Sea, and (i–l) the eastern Pacific.

Beyond reproducing the statistical properties of raw SST, the synthetic SST also captures, in a broad sense, the patterns of key MHW metrics—specifically frequency and duration—although not perfectly (Fig. 4g–l). MHWs are identified by first removing the “true” seasonal mean from both observed and synthetic SST, and then computing a seasonally varying 90th-percentile threshold to define MHW occurrences. This procedure is similar to the two corrected methods, but without the use of moving windows. MHW events are subsequently defined as consecutive days exceeding the threshold; frequency denotes the number of such events, and duration the length (in days) of each event. The results indicate that MHWs derived from synthetic SST occur more frequently but have shorter durations than those derived from observations (Fig. 4h,k). This bias suggests that the synthetic time series still lacks sufficient temporal persistence to sustain prolonged warming events, reflecting missing memory in the representation of SST variability. Nevertheless, the skill of MHW metrics derived from synthetic SST remains substantially higher than that obtained from autoregressive models (Fig.

4i,j), which are commonly used to generate large ensembles of MHW statistics in previous studies. Although not shown here, this advantage is expected to be even more pronounced when considering modulation by large-scale climate modes (e.g., variations across ENSO phases), as such spatial coherence is absent in autoregressive-generated time series.

5. Optimal window size for MHWs

Using the 5,000-year synthetic SST ensembles, we constructed a robust, daily varying long-term climatology, with no smoothing window, that served as the reference (“true”) state. This dataset enables systematic evaluation of the role of moving windows in estimating realistic climatologies from shorter time series with smoothing. We computed the RMSE between the true seasonal mean (from the 5,000-year simulations) and the estimated seasonal mean derived from a limited observational record (41 years) under different moving-window sizes: no window (1 day), 11 days, 31 days, and 61 days (Fig. 6). Results showed that applying a moving window with an appropriate width improved the reconstruction of seasonal climatology, as indicated by lower globally averaged RMSE in most seasons when the window size increased from 1 to 31 days. In contrast, a 61-day window introduced systematic biases relative to the true seasonal mean, yielding RMSE values that exceed those obtained even without a moving window in certain months. These findings indicate that overly large window sizes degrade performance. Together, the results demonstrated that selecting an appropriate window size is critical for accurately estimating baselines for MHW detection. Importantly, they further support the use of the B-corrected method—which incorporates a moving window in the

estimation of seasonal means—over the BV-corrected method when data records are limited (see Fig. 1).

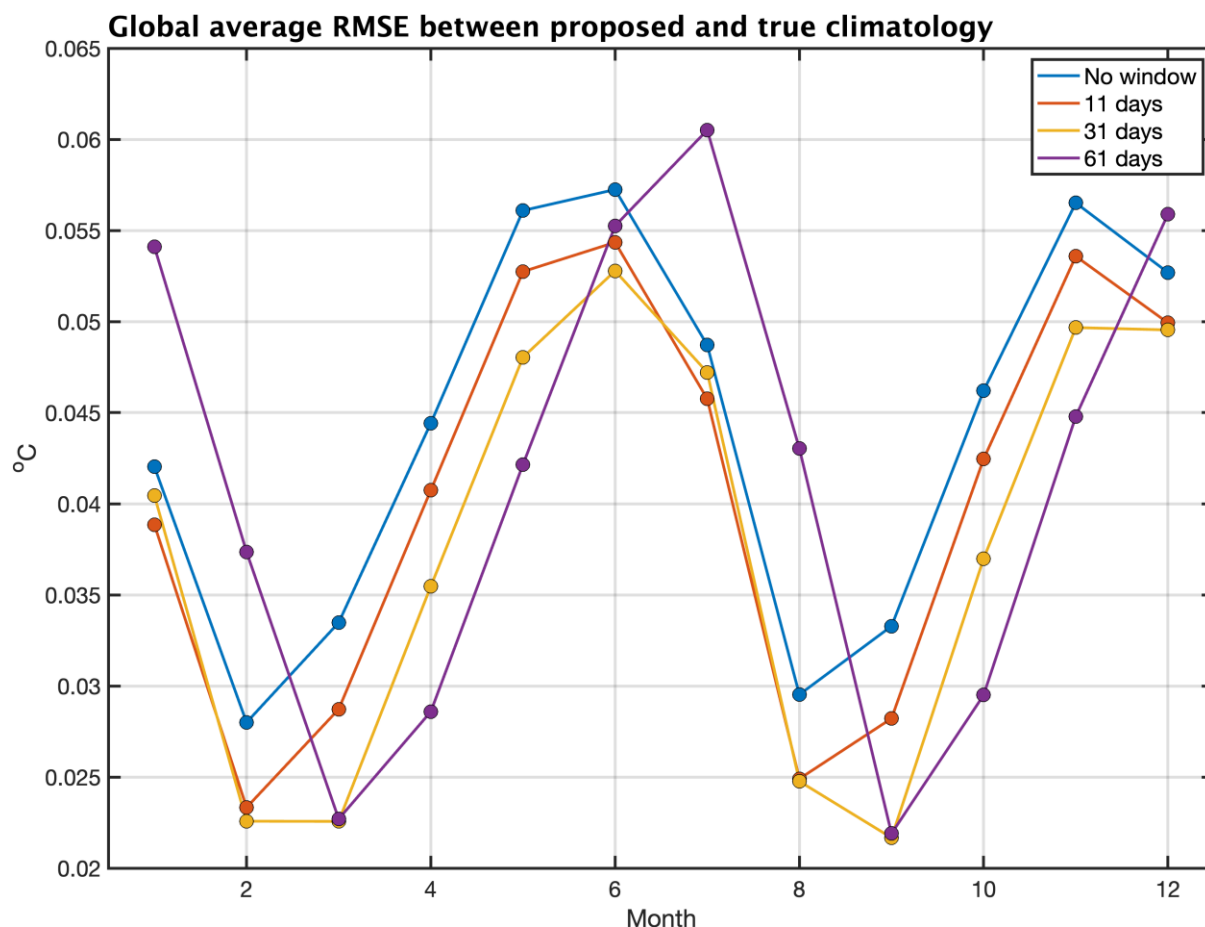


Fig. 6. Global-mean RMSE of seasonal-mean estimates between the true climatology and the proposed climatology under different moving-window lengths. Each line corresponds to a specific window length: no smoothing (1-day window), 11 days, 31 days, or 61 days.

To determine the optimal window size for MHW detection under limited data availability, we computed the MHW severity index from synthetic global SST using baselines derived from varying record lengths—specifically the first 15, 30, and 45 years, typical record lengths for high-resolution model simulations, reanalysis products, and global satellite observations, respectively—and moving window sizes ranging from 1 to 90 days. These baselines were calculated using the B-Corrected method, with moving windows applied to both the seasonal mean and the 90th-percentile threshold. These estimates were then compared with the “true” MHW severity index derived using the full-record seasonal mean and 90th-percentile thresholds. For each record length, the optimal window size was defined as the window that

minimized the RMSE between the true and proposed MHW severity index (Fig. 7). As record length increased, RMSE decreased globally and the optimal window size systematically shortens (Fig. 7a-c). Nevertheless, several regions required optimal window sizes substantially longer than those commonly adopted in temperature extreme definitions (e.g. 11-, 15-, or 31-day windows), including the Indian Ocean, the central-to-western Pacific, the Southern Ocean, and the North Atlantic (Fig. 7d). These relatively long optimal window sizes likely reflect the influence of strong low-frequency and intraseasonal variability (e.g., the Madden-Julian oscillation) and enhanced ocean–atmosphere coupling in these regions. Such variability can obscure the seasonal cycle in short records, leading to a need for stronger temporal smoothing to approximate the underlying climatology.

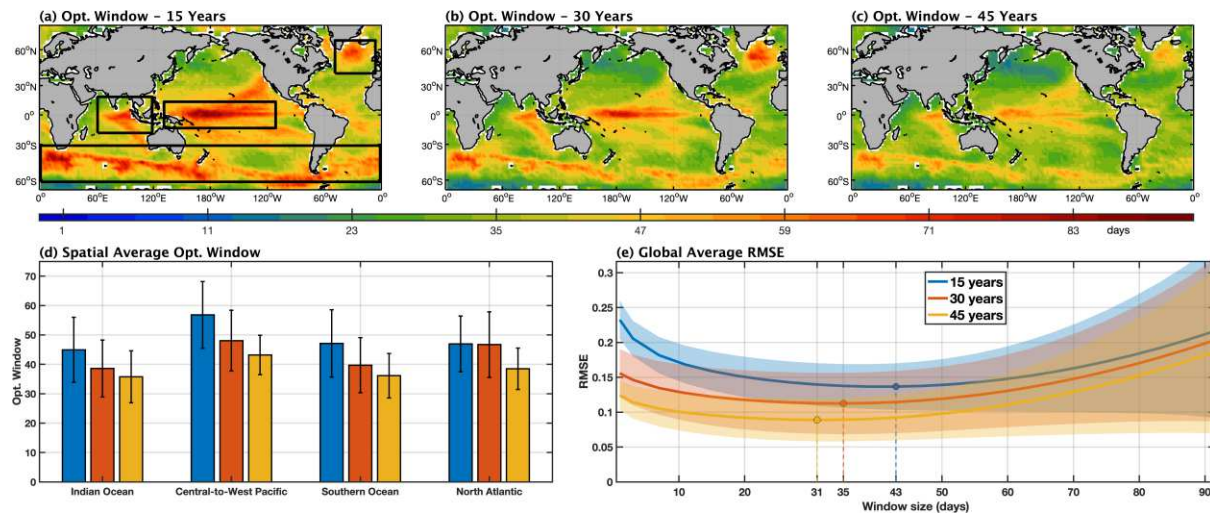


Fig. 7. Optimal window sizes for marine heatwave detection. a–c, Spatial distribution of optimal window sizes determined for each global grid cell using limited record lengths of 15 (a), 30 (b), and 45 years (c). d, Spatially averaged optimal window sizes (mean ± s.d.) for four regions—the Indian Ocean, Central-to-West Pacific, Southern Ocean, and North Atlantic (outlined by black boxes in a)—where systematically larger window sizes are required. e, Global mean RMSE between “true” and “proposed” MHW severity as a function of window size (1–90 days) estimated from limited-length records. Dots denote the optimal window sizes corresponding to the minimum RMSE. Shading indicates spatial variability, estimated as ± 1 spatial standard deviation.

Although the optimal window size varied substantially across spatial scales, providing a single recommended window remains valuable, because a consensus window will facilitate comparability across MHW studies. To this end, we computed the global mean RMSE across all combinations of record lengths and window sizes and identified the optimal window size

for each record length as the one that minimized the global mean RMSE (Fig. 7e). The optimal window size decreased as the length of the record used to construct the baseline increased, consistent with the spatial patterns shown in the maps. For daily datasets spanning 15–45 years, the recommended window sizes ranged from 31 to 43 days (43, 35, and 31 days for 15-, 30-, and 45-year records, respectively). Notably, the 31-day window derived for the 45-year case provides a practical choice for most MHW studies, as many satellite-based SST products, that are frequently used in MHW studies, extend from the early 1980's to the present (~45 years).

In addition to the optimal window sizes derived from MHW severity, optimal window sizes can also be determined for the two seasonal baselines, i.e., the seasonal mean and 90th percentile, by evaluating the RMSE between the true and estimated baselines. Results based on seasonal baselines (seasonal mean, Fig. 8 and seasonal 90th percentile, Fig. 9) are broadly consistent with those obtained using the MHW severity index. The spatial distributions of optimal window sizes derived from seasonal baselines closely resembled results based on MHW severity, with relatively large optimal windows persisting in the Indian Ocean, the central-to-western Pacific, the Southern Ocean, and the North Atlantic. Based on the global mean RMSE, the recommended optimal window sizes are 39 (47), 31 (39), and 31 (35) days for 15-, 30-, and 45-year records, respectively, when using the seasonal mean (90th percentile). Notably, the optimal windows inferred from the MHW severity index was generally between those derived from the two seasonal baselines, suggesting that the severity index integrates the

effects of the moving window on both baseline components in modulating MHW intensity and occurrence.

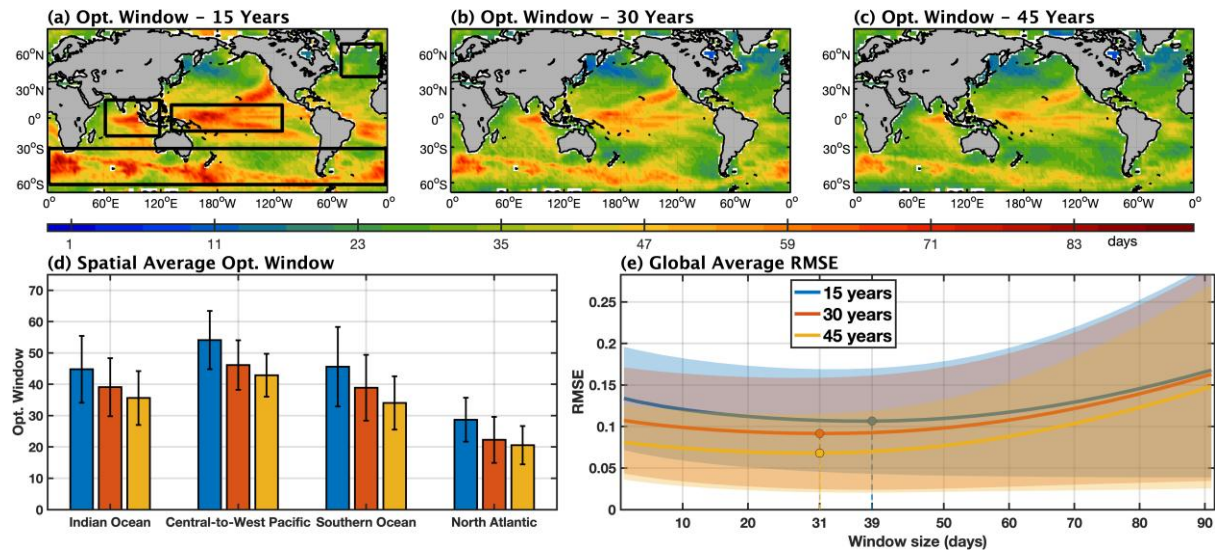


Fig.8. Optimal window sizes for seasonally varying mean. Similar to Fig. 7, but with results estimated from seasonally varying mean.

Although the corrected methods have been shown to provide a more robust framework for identifying MHWs, the uncorrected method remains widely used in studies of extremes and may continue to be adopted for simplicity and consistency with previous studies. Therefore, offering simple and easy-to-implement guidance to identify the optimal window sizes for the uncorrected method remains valuable for future MHW analyses. To this end, we computed both the proposed and true MHW severity indices using seasonal baselines derived from the uncorrected method, where seasonal mean and 90th percentile are calculated independently from the raw synthetic time series (Fig. 1). We then evaluated optimal window sizes across a range of window lengths and record length (Fig. 10). Global-mean RMSE indicated that the optimal window sizes are 35, 27, and 19 days for 15-, 30-, and 45-year records, respectively. These values were substantially smaller than those recommended for the B-corrected method (Fig. 7e), yet are still larger than the 11-day window adopted in the standard Hobday et al. MHW definition.

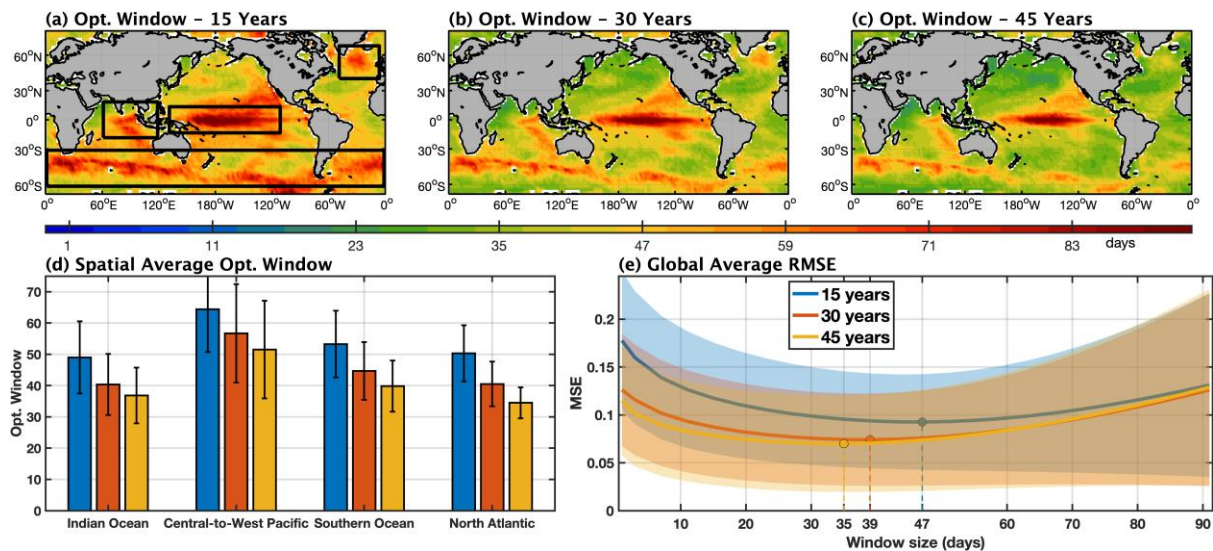


Fig. 9. Optimal window sizes for seasonally varying 90th percentile. Similar to Fig. 7, but with results estimated from seasonally varying 90th percentile.

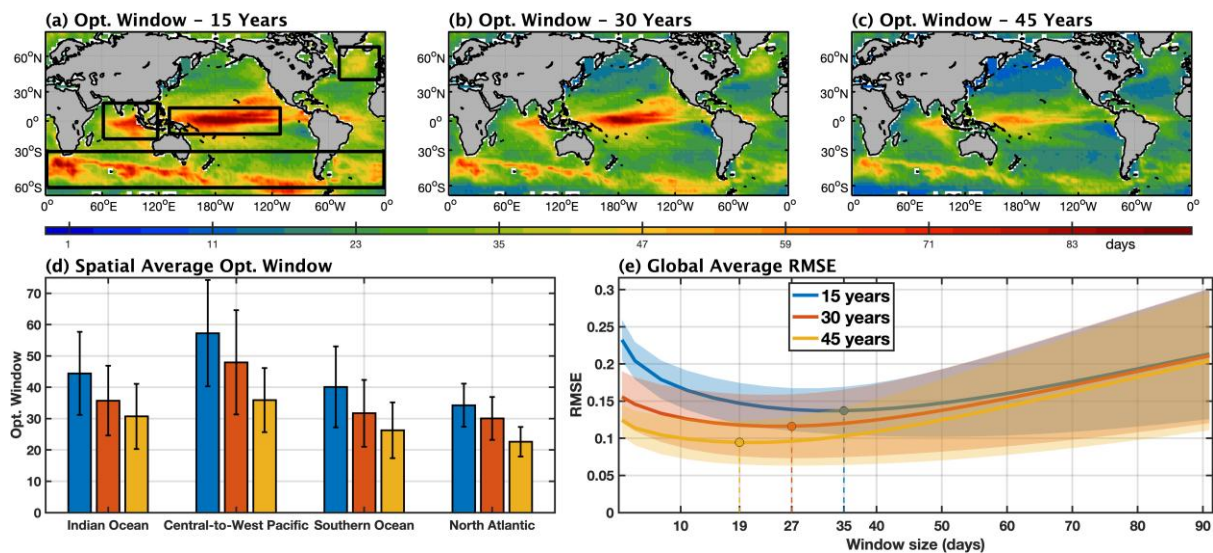


Fig. 10. Optimal window sizes for uncorrected approach. Similar to Fig. 7, but with results estimated from MHW severity index estimated from uncorrected approach.

6. Practical consequences of detection-method choice

Finally, we quantify how much the choice of detection method actually alters our understanding of MHWs, by comparing global observational estimates obtained from the Hobday definition against the corrected method using the optimal moving window for each timescale (43-, 35-, and 31-day windows for 15-, 30-, and 45-year records, respectively; Fig. 7). We applied both methods to OISST using baselines constructed over two 15-year periods (1993–2007 and 2008–2022), a 30-year period (1982–2012), and an approximately 45-year

period (1982–2025), across a range of global average MHW metrics (Fig. 11), quantifying the discrepancy between methods by the mean absolute percentage error (MAPE) between the resulting time series.

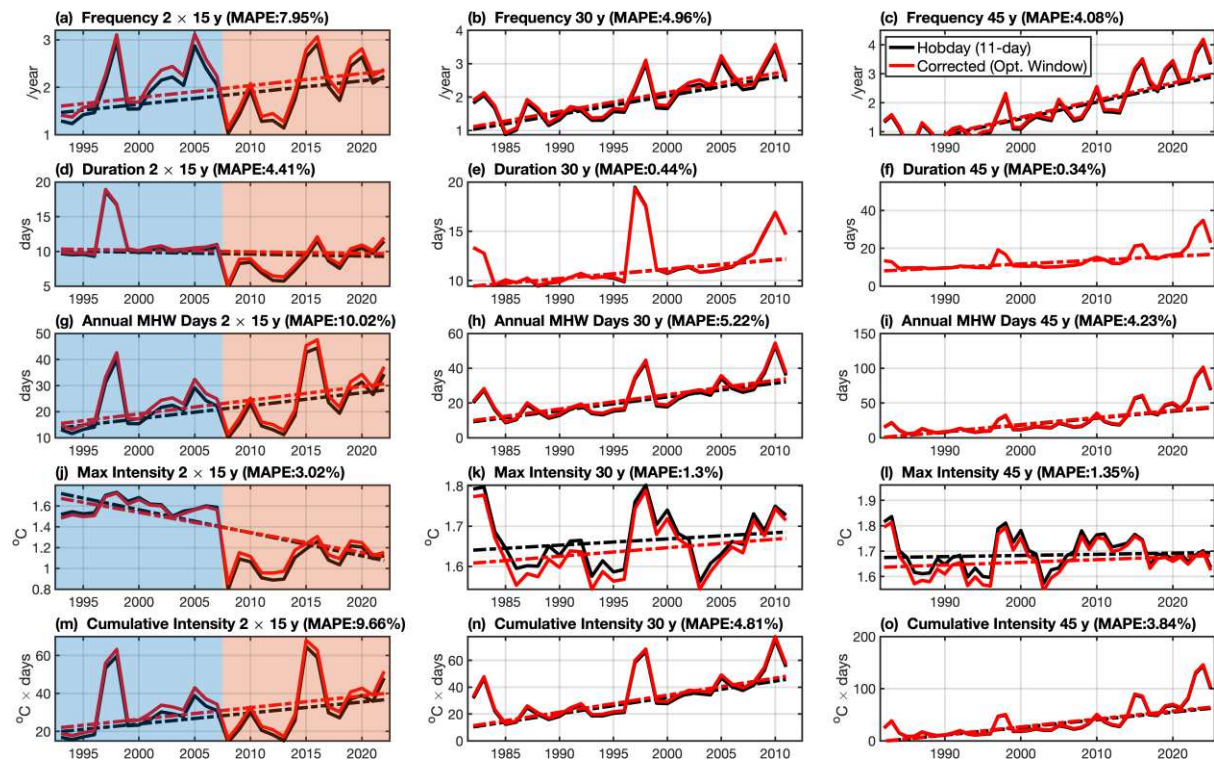


Fig. 11. Global marine heatwave metrics from conventional and optimized detection methods. Global average annual MHW metrics were calculated from OISST using the Hobday definition (black) and the corrected method with the optimal moving-window sizes identified in Fig. 7 (red). Metrics include (a–c) frequency, (d–f) mean duration, (g–i) annual MHW days, (j–l) annual maximum intensity, and (m–o) annual cumulative intensity. Baselines were constructed using two 15-year periods (1993–2007 and 2008–2022), one 30-year period (1982–2012), and an approximately 45-year period (1982–2025). Panel titles report the mean absolute percentage error (MAPE) between the two methods.

Two features stand out. First, the discrepancy is metric-dependent: duration and maximum intensity are relatively insensitive to the choice of method, whereas frequency, annual MHW days, and annual cumulative intensity are substantially more affected. Second, and more consequentially, the discrepancy grows at an accelerating rate as the baseline shortens (the increase from the 30- to the 15-year baseline exceeds that from the 45- to the 30-year baseline) reaching 10% for the most affected metrics with only 15 years of data. This behaviour implies

that methodological artefacts remain modest for the multidecadal (30–45 year) baselines typical of observational studies, but escalate rapidly as records shorten.

7. Discussion

This study addressed a fundamental yet underexplored problem in the detection of temperature extremes: how to define extremes objectively when observational records are limited and noisy. By combining a bias-aware detection method with a novel synthetic SST framework for optimizing baseline parameters, specifically the moving-window size, we provide an optimized framework for detecting temperature extremes that represents both a conceptual advance in extreme-event definition and a practical solution to a long-standing methodological challenge.

A central contribution of this work is the development of a new framework for generating large ensembles of synthetic global SST fields that accurately reproduce the key statistical properties of the observed ocean surface. Unlike previous approaches that rely on simplified stochastic models (Brunner; Voigt 2024; Oliver 2019; Wang et al. 2022; Xu et al. 2022) or computationally expensive climate model ensembles (Dong et al. 2024) that may still exhibit significant biases, our method preserves the observed seasonal mean and variance structure, daily-to-subseasonal persistence, and large-scale climate modes, within a unified statistical framework. Importantly, synthetic time series are constrained by observations rather than by model physics, allowing the generated SST realizations to remain statistically consistent with real world data while extending far beyond the limited length of the satellite record. To our knowledge, previous observation-based synthetic time series approaches have either relied on local stochastic models that neglect large-scale climate-mode coupling, or on resampling techniques that cannot generate statistically independent long-term realizations beyond the observational record (Baugh; McKinnon 2022; McKinnon et al. 2017). Here we demonstrate instead the ability to generate synthetic SST time series that simultaneously retain local autocorrelation, remote climate-mode influence, and realistic spatiotemporal variability at the global scale. This capability enables the construction of a statistically robust “synthetic truth” against which extreme-event detection methodologies can be objectively evaluated.

Leveraging this synthetic framework, we address a long-standing issue in MHW research: the absence of an objective criterion for choosing the moving-window length used to define climatological baselines. Previous studies have largely adopted window sizes (e.g., 11, 15, or

31 days) based on convention or analogy with atmospheric heatwaves, without a quantitative justification. Our results demonstrate that window choice can strongly influence both the detected frequency and intensity of MHWs, particularly when data records are short. By defining optimality in terms of minimizing the error in the MHW severity index relative to a long-record reference, we provide the first statistically grounded determination of an optimal window size. This approach moves the definition of temperature extremes away from heuristic choices and toward an explicitly data-informed, reproducible standard.

The divergent temporal evolution of MHWs identified by the loose Hobday framework and the optimal detection framework raises particular concerns for analyses based on records substantially shorter than the 30-year standard. Although global observational and reanalysis products typically span 30–50 years, a non-negligible number of studies are constrained to considerably shorter baselines (in some cases around 15 years or less) owing to the high computational cost of long, high-resolution simulations with comprehensive diagnostics (Bian et al. 2024; Bian et al. 2023), or the scarcity of long, high-quality subsurface and fine-scale observations (Ma; Chen 2025; Pinter et al. 2026; Plume et al. 2025). Such records fall precisely within the range where we find methodological artefacts to escalate. For analyses constrained by these limited records, applying the optimal framework can substantially reduce method-dependent biases; we nonetheless emphasize that the intrinsic uncertainty of short baselines cannot be fully removed by any detection method, and longer records should still be prioritized where feasible.

Beyond MHWs, the methodological implications of this work are considerably broader. The combination of synthetic data generation with objective optimization of baseline definitions offers a general framework for defining mean or threshold climatologies in noisy, short observational records. Similar challenges arise in the detection of other types of weather and climate extremes that rely on percentile-based thresholds derived from limited data. The analysis pathway developed here—constructing a statistically accurate synthetic reference, quantifying detection error, and optimizing methodological choices accordingly—can be readily extended to different types of extreme events.

In summary, this work advances the study of temperature extremes in two key ways: it introduces an observation-constrained framework for generating realistic long-term synthetic records, and it provides the first objective, quantitative determination of optimal baselines, including recommended moving-window sizes. As climate-driven extremes continue to

intensify and expand, such statistically robust and transparent approaches are essential to ensure that detected trends reflect physical change rather than methodological artefacts.

Acknowledgments.

This work is an output of the International Marine Heatwaves Working Group (www.marineheatwaves.org). L.B. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2037 'CLICCS—Climate, Climatic Change, and Society' - Project No. 390683824, a contribution to the 'Earth and Society Research Hub' (ESRAH) of University of Hamburg. K.F.D acknowledges funding support from the Australian Research Council (FT230100214). E.C.J.O acknowledges funding support from the National Science and Engineering Research Council (RGPIN2024-04800). N.J.H acknowledges funding support from the ARC Centre of Excellence for the Weather of the 21st Century (CE230100012) and ARC Centre of Excellence for Climate Extremes (CE170100023). A.S.G acknowledges funding support from the Australian Research Council (FT220100475). D.A.S was supported by a UKRI Future Leaders Fellowship (MR/X023214/1). M.S.T was supported by the New Zealand Ministry of Business, Innovation and Employment (Toka ākau toitu Kaitiakitanga – building a sustainable future for coastal reef ecosystems). T.W. acknowledges funding support from the Australian Research Council (FL240100015). K.E.S was supported by funding from the Challenger Society for Marine Science. This work was supported by computational resources provided by the Australian Government through the National Computational Infrastructure (NCI) under the National Computational Merit Allocation Scheme and ANU Merit Allocation Scheme. The authors declare no conflict of interests.

Data Availability Statement.

The NOAA OI SST used in this study can be obtained via <https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.highres.html>. For this study, the datasets were retrieved and stored on Gadi, the supercomputing facility operated by Australia's National Computational Infrastructure. Codes used to generate observation-constrained synthetic time series can be found via <https://github.com/ZijieZhaoMMHW/SynTS..>

APPENDIX

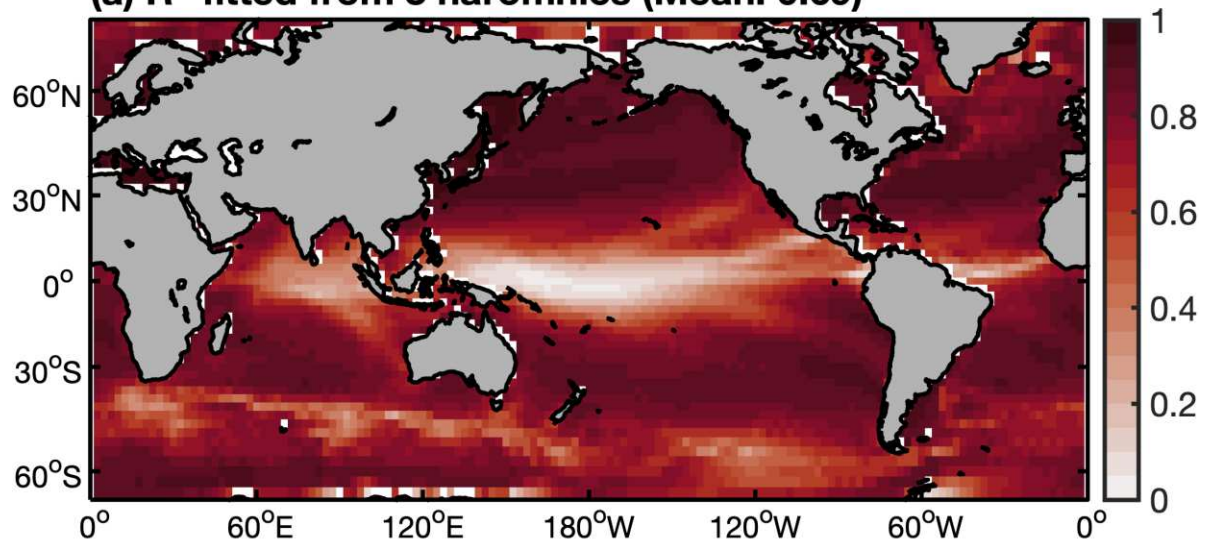
Appendix A: Determination of optimal harmonic climatology

Compared to estimate climatology using prescribed orders of harmonics, a more objective way may be to use a stepwise process. A sequence of harmonic regression can be fitted to a given time series as following the formula:

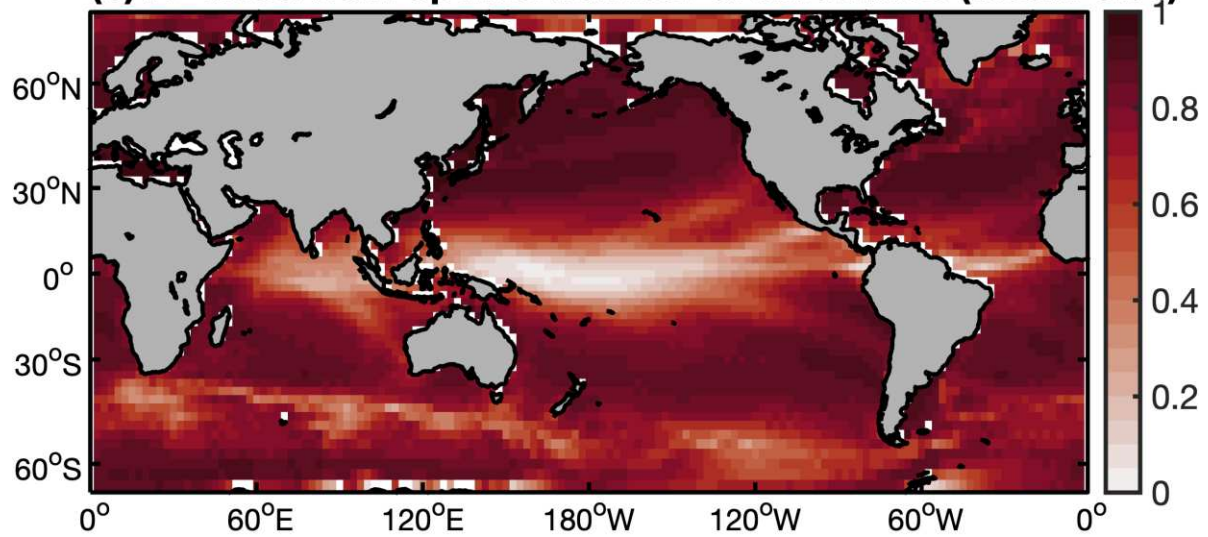
$$\mu_j(d(t)) = c + \sum_{i=1}^j \left[a_i \cos \left(\frac{2\pi i d(t)}{365} \right) + b_i \sin \left(\frac{2\pi i d(t)}{365} \right) \right],$$

where $d(t)$ is the calendar day at time t (1979-2019). At each step, the model order is increased by one only if at least one of the two newly added coefficients, a_i or b_i , is significant at the 95% confidence level, after which the regression is refitted. The procedure stops when neither newly added coefficient is significant; the harmonic order retained from the preceding step is then taken as the optimal order. Here we compare the goodness of fit, measured by R^2 , between the optimal harmonic model and the third-order harmonic model used in this study (Fig. A1), suggesting that the two methods demonstrate similar fitting skills when measuring climatology.

(a) R^2 fitted from 3 haromnics (Mean: 0.69)



(b) R^2 fitted from optimal number of haromnics (Mean: 0.69)



(c) Optimal number of harmonics for (b)

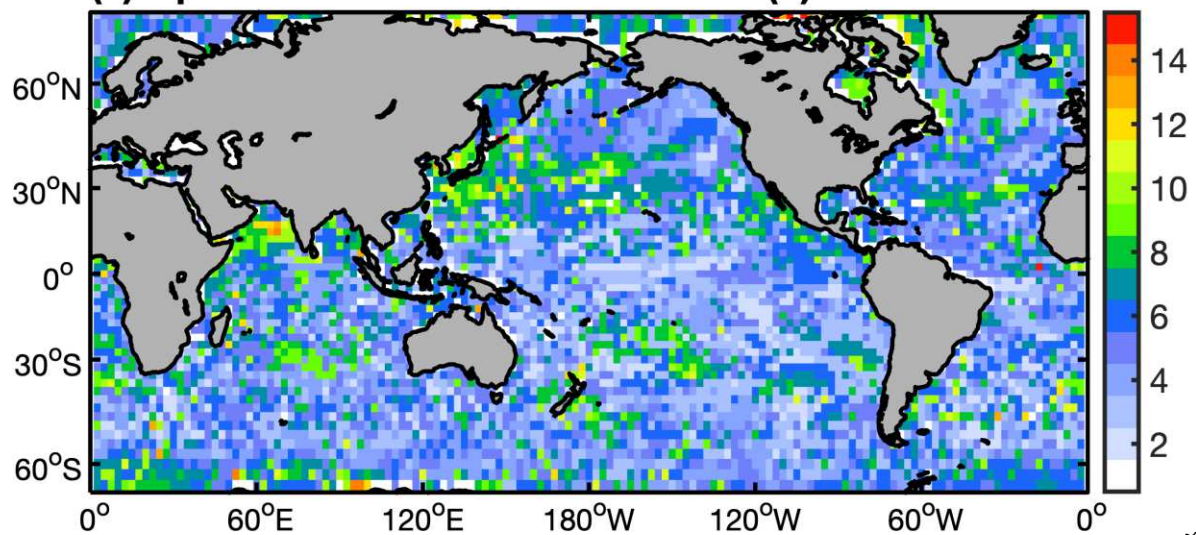


Fig. A1. Performance of the harmonic climatology. (a,b) R^2 from harmonic regressions fitted to SST using (a) three harmonics and (b) the optimal number of harmonics. The optimal number was determined by stepwise regression, in which harmonics were added sequentially until the newly added terms were no longer statistically significant at the 95% confidence level. The global mean R^2 is reported in the title of each panel. (c) Optimal number of harmonics used in b.

Appendix B: Rationale of PC numbers when fitting LIMs

Using the method described in section 2, we generated 1000-member ensembles of 5000-yr synthetic SST time series constrained by observations from 1982 to 2022, using either 20 or 100 principal components. The simulations based on 20 PCs better reproduce the observed lag-autocorrelation structure, whereas those based on 100 PCs tend to overestimate autocorrelation. We attribute this discrepancy to the monthly formulation of the LIM, which can inflate persistence when the model is applied to daily-to-monthly variability.

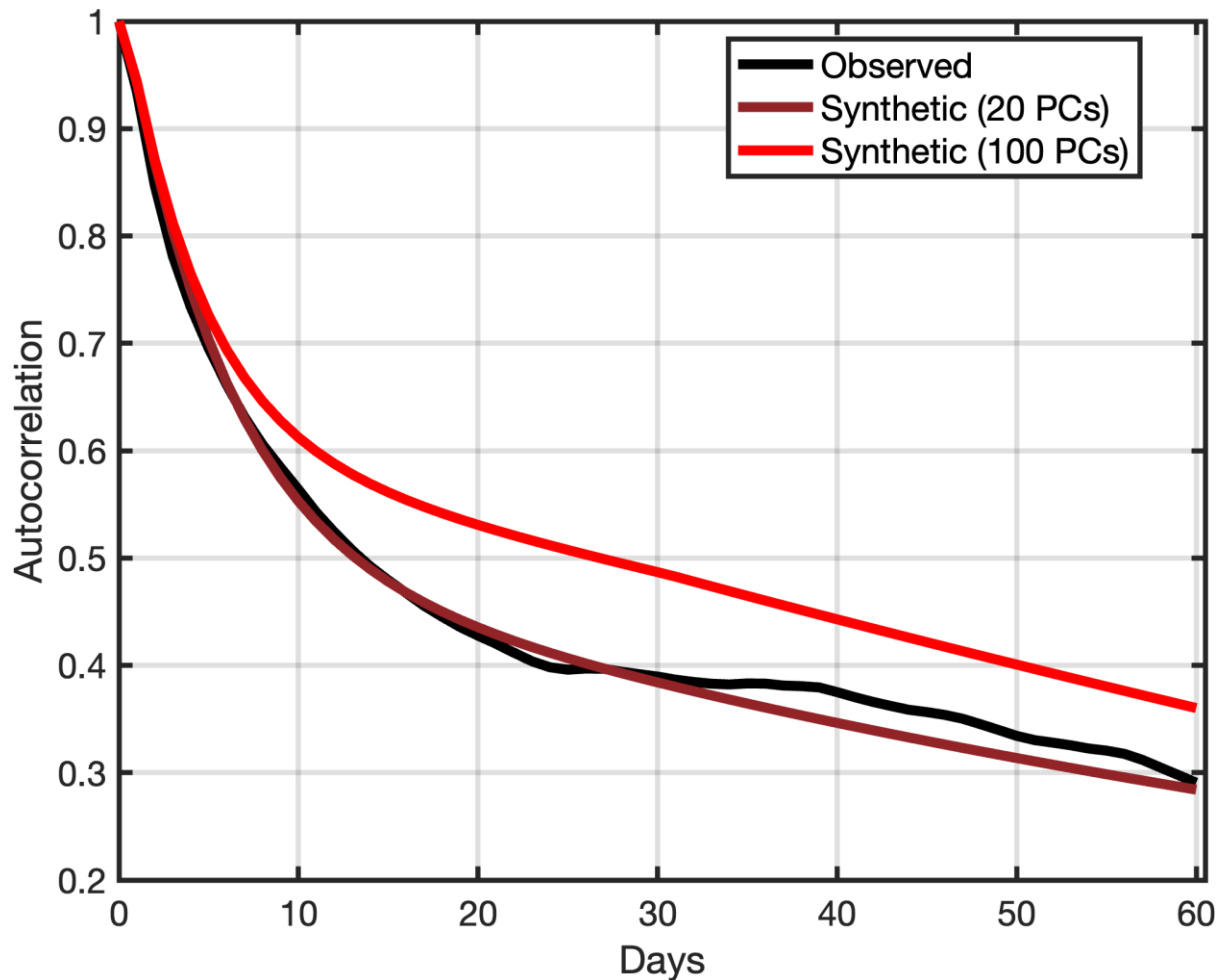


Fig. B1. Effect of the number of principal components in the Linear Inverse Model on the autocorrelation of synthetic SST anomalies. As in Fig. 4d, but showing autocorrelation functions at lags of 1–60 days for observed SST anomalies and synthetic SST anomalies with remote components constructed from LIMs using 20 or 100 PCs.

Appendix C: Sensitivity of optimal window size to long-term trends

To examine if the determined optimal window size is sensitive to the enrollment of long-term trend in the synthetic SST time series, we did a sensitivity test (Fig. C1). We first simulated 1000 ensembles of 5000-year synthetic SST constrained by observed SST record at 106.25° E, 8.75° S during 1982–2022 following the method described in Section 2. The seasonal mean and 90th percentile calculated from the full synthetic time series were treated as the reference climatology. We then extracted the first 45 years of the synthetic record, imposed zero-mean linear trends with varying slopes, and estimated the seasonal mean and 90th percentile using window sizes of different lengths. For each window size and trend slope, the RMSE was calculated between the MHW severity indices derived from the reference and estimated climatologies. Across all imposed trends, the minimum RMSE was consistently obtained with a window size of 35 days.

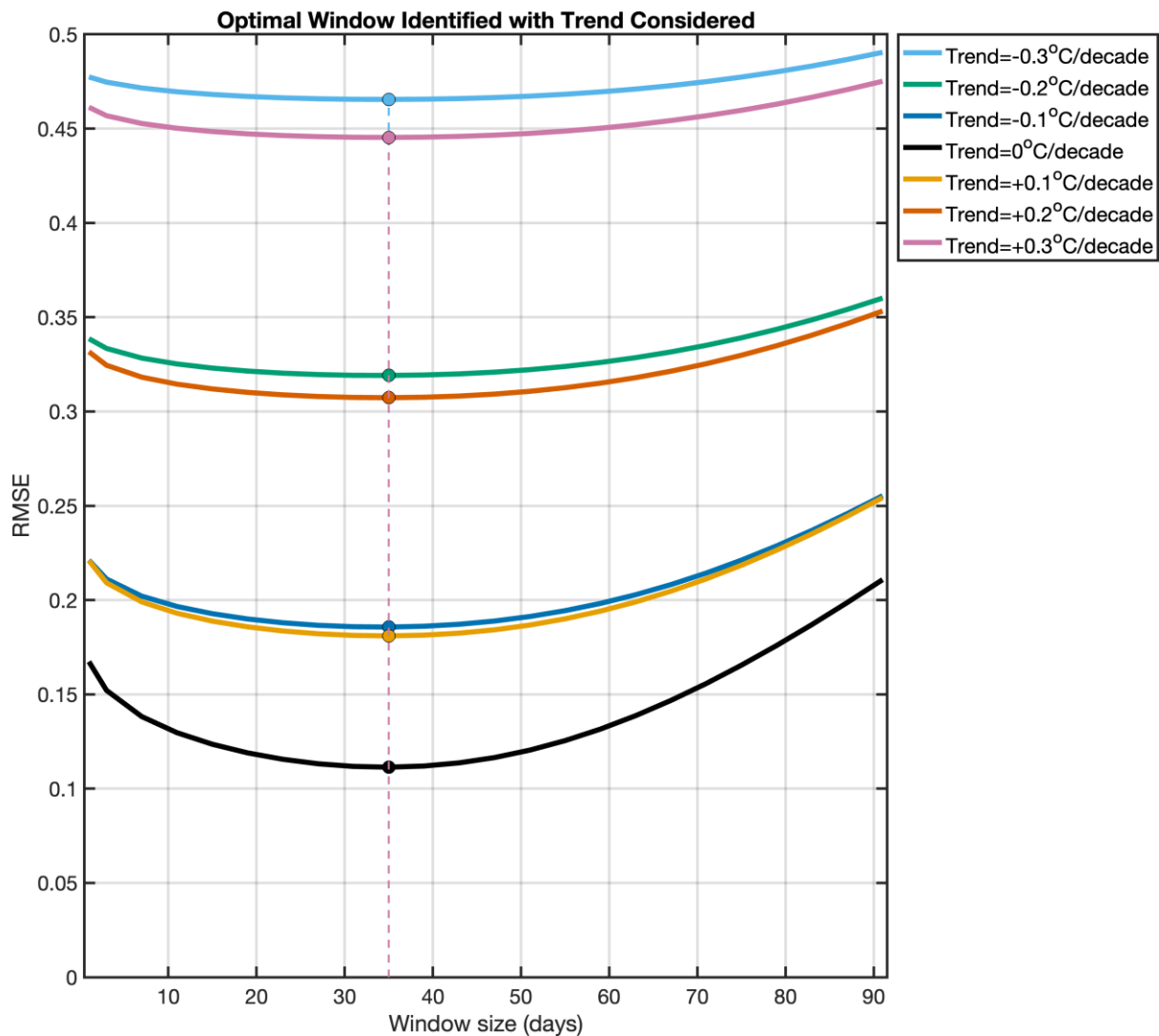


Fig. C1. Sensitivity of the estimated optimal window size to imposed trends. The analysis is based on 1000 ensemble members of 5000-yr synthetic SST time series constrained by observed SST at 106.25°E, 8.75°S. Linear trends ranging from -0.3 to $0.3^{\circ}\text{C decade}^{-1}$ were imposed on the synthetic time series. The estimated optimal window size remains constant at 35 days across all imposed trends.

REFERENCES

- Amaya, D. J., and Coauthors, 2023: Marine heatwaves need clear definitions so coastal communities can adapt. *Nature*, **616**, 29-32.
- Baugh, S., and K. McKinnon, 2022: Hierarchical Bayesian modeling of ocean heat content and its uncertainty. *The Annals of Applied Statistics*, **16**, 2603-2625.
- Bian, C., Z. Jing, H. Wang, and L. Wu, 2024: Scale-dependent drivers of marine heatwaves globally. *Geophysical Research Letters*, **51**, e2023GL107306.
- Bian, C., Z. Jing, H. Wang, L. Wu, Z. Chen, B. Gan, and H. Yang, 2023: Oceanic mesoscale eddies as crucial drivers of global marine heatwaves. *Nature communications*, **14**, 2970.

Brunner, L., and A. Voigt, 2024: Pitfalls in diagnosing temperature extremes. *Nature Communications*, **15**, 2087.

Dong, L., S. Zeng, F. Song, L. Wu, J. Shi, Y. Ding, and S. Lv, 2024: Roles of external forcing and internal variability in global marine heatwaves change during 1982–2021. *Geophysical Research Letters*, **51**, e2023GL107557.

Freychet, N., G. Hegerl, D. Mitchell, and M. Collins, 2021: Future changes in the frequency of temperature extremes may be underestimated in tropical and subtropical regions. *Communications Earth & Environment*, **2**, 28.

Hobday, A. J., and Coauthors, 2016: A hierarchical approach to defining marine heatwaves. *Progress in Oceanography*, **141**, 227–238.

Levitus, S., 1987: A comparison of the annual cycle of two sea surface temperature climatologies of the world ocean. *Journal of Physical Oceanography*, **17**, 197–214.

Li, C., J. Huang, X. Liu, L. Ding, Y. He, and Y. Xie, 2024: The ocean losing its breath under the heatwaves. *Nature Communications*, **15**, 6840.

Liu, H., X. Nie, C. Cui, and Z. Wei, 2023a: Compound marine heatwaves and low sea surface salinity extremes over the tropical Pacific Ocean. *Environmental Research Letters*, **18**, 064001.

Liu, H., T. Xu, D. Li, X. Nie, and Z. Wei, 2023b: Sea surface salinity extremes over the global ocean. *Environmental Research Letters*, **18**, 114020.

Ma, X., and G. Chen, 2025: Marine heatwaves are shaping the vertical structure of phytoplankton in the global ocean. *Communications Earth & Environment*, **6**, 715.

McKinnon, K. A., A. Poppick, E. Dunn-Sigouin, and C. Deser, 2017: An “observational large ensemble” to compare observed and modeled temperature trend uncertainty due to internal variability. *Journal of Climate*, **30**, 7585–7598.

Oliver, E. C., 2019: Mean warming not variability drives marine heatwave trends. *Climate Dynamics*, **53**, 1653–1659.

Pearce, A. F., R. Lenanton, G. Jackson, J. Moore, M. Feng, and D. Gaughan, 2011: *The "marine heat wave" off Western Australia during the summer of 2010/11*. Western Australian Fisheries and Marine Research Laboratories Hillarys, WA

Penland, C., and P. D. Sardeshmukh, 1995: The optimal growth of tropical sea surface temperature anomalies. *Journal of climate*, **8**, 1999–2024.

Perkins, S. E., and L. V. Alexander, 2013: On the measurement of heat waves. *Journal of climate*, **26**, 4500–4517.

Pinter, S. S., N. L. Jones, M. D. Rayson, and M. V. Cuttler, 2026: A deeper look into marine heatwaves off the coast of Western Australia. *Weather and Climate Extremes*, 100906.

Plume, G., and Coauthors, 2025: Vertical structure of subsurface marine heatwaves in a shallow nearshore upwelling system. *Scientific Reports*, **15**, 6353.

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

Russo, S., J. Sillmann, and E. M. Fischer, 2015: Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters*, **10**, 124003.

Schlegel, R. W., E. C. Oliver, A. J. Hobday, and A. J. Smit, 2019: Detecting marine heatwaves with sub-optimal data. *Frontiers in Marine Science*, **6**, 737.

Schlegel, R. W., S. Darmaraki, J. A. Benthuisen, K. Filbee-Dexter, and E. C. Oliver, 2021: Marine cold-spells. *Progress in Oceanography*, **198**, 102684.

Sen Gupta, A., and Coauthors, 2020: Drivers and impacts of the most extreme marine heatwave events. *Scientific reports*, **10**, 19359.

Smale, D. A., and Coauthors, 2019: Marine heatwaves threaten global biodiversity and the provision of ecosystem services. *Nature Climate Change*, **9**, 306-312.

Smith, K. E., and Coauthors, 2021: Socioeconomic impacts of marine heatwaves: Global issues and opportunities. *Science*, **374**, eabj3593.

Smith, K. E., and Coauthors, 2023: Biological impacts of marine heatwaves. *Annual review of marine science*, **15**, 119-145.

Smith, K. E., and Coauthors, 2024: Global impacts of marine heatwaves on coastal foundation species. *Nature Communications*, **15**, 5052.

Smith, K. E., and Coauthors, 2025: Baseline matters: Challenges and implications of different marine heatwave baselines. *Progress in Oceanography*, **231**, 103404.

Thoral, F., and Coauthors, 2026: Marine darkwave as an event-based framework to assess unusual periods of reduced underwater light availability. *Communications Earth & Environment*, **7**, 4.

Vogel, M. M., J. Zscheischler, E. M. Fischer, and S. I. Seneviratne, 2020: Development of future heatwaves for different hazard thresholds. *Journal of Geophysical Research: Atmospheres*, **125**, e2019JD032070.

Wang, Y., J. B. Kajtar, L. V. Alexander, G. S. Pilo, and N. J. Holbrook, 2022: Understanding the changing nature of marine cold-spells. *Geophysical Research Letters*, **49**, e2021GL097002.

Wernberg, T., and Coauthors, 2025: Marine heatwaves as hot spots of climate change and impacts on biodiversity and ecosystem services. *Nature Reviews Biodiversity*, 1-19.

Woolway, R. I., E. Jennings, T. Shatwell, M. Golub, D. C. Pierson, and S. C. Maberly, 2021: Lake heatwaves under climate change. *Nature*, **589**, 402-407.

Xu, T., M. Newman, A. Capotondi, S. Stevenson, E. Di Lorenzo, and M. A. Alexander, 2022: An increase in marine heatwaves without significant changes in surface ocean temperature variability. *Nature Communications*, **13**, 7396.

Zhang, X., and Coauthors, 2011: Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change*, **2**, 851-870.

Zscheischler, J., and Coauthors, 2020: A typology of compound weather and climate events. *Nature reviews earth & environment*, **1**, 333-347.