Supporting Information for "On the vertical structure and propagation of marine heatwaves in the Eastern Pacific"

Eike E. Köhn[1], Meike Vogt[1], Matthias Münnich[1], Nicolas Gruber[1,2]

[1]Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, 8092 Zurich, Switzerland

[2]Center for Climate Systems Modeling, ETH Zurich, 8092 Zurich, Switzerland

Corresponding author: E. E. Köhn, Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, 8092 Zurich, Switzerland (eike.koehn@usys.ethz.ch)

May 25, 2023, 11:09am

Contents of this file

1. Model data description and evaluation

1.1. Vertical regridding of model output

To account for the bathymetry following native model grid, we perform vertical regridding before calculating horizontal averages when horizontally coarsening/downsampling the model output (Sec. 2.1 of main text). We therefore regrid the model output from the bathymetry following s-level coordinates to z-levels, that is fixed depth levels, within the upper 500 m. We thereby choose the the z-levels to closely follow the vertical spacing in the maximally stretched s-level coordinate system. We therefore let the z-levels closely follow the s-levels at a location of the deepest model bathymetry, which is set to 6500 m (Fig. S2). This leads to 37 z-levels with a 5 m resolution in the upper 100 m and a gradual increases to a 50 m resolution below 350 m. By fitting the z-levels to the maximally stretched s-level coordinates, the regridding to z-levels does not add additional intermediate depth levels. Vertical temperature profiles from locations where the ocean bottom is shallower than 6500 m have a higher resolution in s-levels than in z-levels in the upper 500 m.

To test for the sensitivity of our results regarding the vertical resolution, we additionally coarsen the vertical z-level grid, by using only every second z-level for the MHW analysis, leading to 19 z-levels (Fig. S2, Section 2.6 of main text).

1.2. Evaluation of the mean state

The hindcast mean sea surface temperature (SST, 1979-2019) has a small bias of −0.02 °C compared to World Ocean Atlas 2018 (WOA2018, Boyer et al. (2018)) averaged across the full study area (Fig. S3). In the equatorial Eastern Pacific (EP) and Peruvian upwelling system, model SST values are locally up to 1.5 °C warmer than observed. The temperature biases are stronger in the subsurface with pronounced regional differences (Fig. S3). As such, at 200 m (400 m) depth, the tropical EP between 20°S and 20°N is on average too cold by 1.47 °C (0.99 °C), while the subpolar North Pacific north of 30°N is too warm by 0.41 °C (1.17 °C). At 100 m, we find additional warm biases in the eastern tropical North Pacific and the subtropical South Pacific with mean anomalies of up to 1.5 °C. This warm bias is linked to a slightly too deep mixed layer (and thermocline) in the subtropical gyres (∼10–30 m, Fig. S4a). Across the entire EP, the modeled mixed layer depth (MLD) is on average 12.4 m deeper than observed (Holte et al., 2017). In the subpolar NP, ROMS simulates a MLD that is 4.5 m too shallow. ROMS accurately reproduces the sea surface height (SSH) field structure (spatial correlation of the temporal mean SSH fields of 0.98 , Fig. S4b), indicating that the mean geostrophic currents are well reproduced by the model.

1.3. Evaluation of the temperature variability

We evaluate the variability of the SST anomalies, by analysing their amplitudes and persistence in ROMS. The anomalies are thereby calculated relative to the seasonally varying SST climatology calculated for the period 1982–2011 (Fig. S5). For the evaluation, we compare the standard deviation and the autocorrelation e-decay time scale of the daily SST anomalies and compare the results to observational OISSTv2 data (Reynolds et al., 2007). The model reproduces the SST anomaly amplitudes reasonably well with maximum standard deviations of up to 1.5 °C in the tropical EP (Fig. S5a,b). In the Peruvian upwelling system the model overestimates the generally high standard deviation by about 0.5 °C. The persistence of SST anomalies, or the time scale of SST anomaly variability is also well reproduced by the model with generally longer e-decay times in the tropical EP, the subtropical NP gyre and the subpolar NP (Fig. S5c,d). However, the model has a tendency to overestimate the e-decay times, indicating longer persistence of SST anomalies. This is in agreement with the detection of generally longer but fewer marine heatwaves (MHWs, Sec. 3 of the main text, Oliver et al. (2021)).

1.4. Evaluation of temperature trends

We calculate trends in simulated the temperature field at the surface, at 100 m, 200 m, 310 m and 400 m depth over the full hindcast period (1979-2019, Fig. S6).

We find a general simulated cooling trend of around $-0.03\,^{\circ}\mathrm{C\,yr^{-1}}$ throughout the tropical EP across all depths (Fig. S6a-e). The cooling trend is most pronounced at 100 m depth (up to $-0.05\,^{\circ}\mathrm{C\,yr^{-1}}$), that is in the strong thermocline of the tropical EP. This cooling trend is stronger than comparable temperature trends calculated from the monthly resolved EN4 data set from 1981–2019 (Good, Martin, and Rayner (2013), Fig. S6f-j). The analysis of temperature time series shows that the calculation of long-term trends in the tropical EP is influenced by interannual temperature variability (Fig. S7). For instance, in the northern Humboldt Current System (location *c* indicated in Figure S6), the occurrence of fewer strong El Niño related warming events after the year 2000 compared to the preceding years, manifests in the diagnosis of a general cooling of the EP.

Below 200 m depth the model shows a pronounced cooling between 25°-40°N and 160°-120°W, which is not found in the EN4 data set from 1981–2019 (Good et al. (2013), Fig. S6c-e,h-j). The temperature time series within this region (chosen location *a*, Figure S6), shows that this cooling trend is relatively independent from interannual variability, but manifests mostly at the beginning of the hindcast with the strongest temperature decrease in the first few years (Fig. S7a). This suggests some model adjustment occurring when switching from the spin-up period to the beginning of the model hindcast. Such spurious trends in the simulated temperature field have the potential to affect the detection of extreme temperatures. We however limit the impact of this spurious model trend by calculating the temperature thresholds based on the years 1982–2011 (Sec. 2.2.1 of main

text) and thus by not taking the first three hindcast years into account. Nevertheless, as the MHW detection is performed over the full analysis time period, the spurious model drift leads locally to higher numbers of subsurface extreme temperatures at the beginning of the hindcast (Fig. S8, Fig. S9). To assess the influence of the first few hindcast years on the composite MHW characteristics, we conduct a sensitivity analysis and calculate the MHW characteristics only for the time period 1982–2019 (case G in Section 4 on sensitivities analyses). The results of the sensitivity analysis however suggest only limited impacts on the statistics of the one-dimensional MHW properties (see Fig. S19).

1.5. Evaluation of subsurface temperature extremes

Only few observations exist, documenting the structure and evolution of MHWs in the subsurface. This complicates an evaluation of the model's capability to realistically simulate subsurface temperature extremes. As we study MHWs in the EP between 1979–2019 we are able to compare our model results to the subsurface evolution of the "Blob" as described by Scannell, Johnson, Thompson, Lyman, and Riser (2020) (see Sec. 4.3 of main text). Furthermore, we can make use of the TAO/TRITON mooring array in the tropical Pacific (McPhaden et al., 2010), which has gathered temperature time series at multiple locations within the upper 500–750 m of the water column. This type of temperature time series have already been used to calculate MHW characteristics in the western tropical Pacific (Hu et al., 2021).

Here we use the available temperature time series from five equatorial TAO/TRITON moorings east of 155°W, to evaluate the models performance in reproducing the vertical structure of temperature anomalies in the equatorial EP. As a test case we therefore focus on the 1997–1998 El Niño event (Fig. S10), which was an intense and long lasting MHW in the tropical EP (Sen Gupta et al., 2020). For the hindcast simulation and the TAO/TRITON array, we calculate temperature anomalies relative to the climatology, which we calculate following Hobday et al. (2016). For the hindcast, the climatology is calculated over the 1982–2011 period, as throughout the main study. For the TAO/TRITON data, we use all available data to calculate the climatologies, as the mooring data generally covers varying time spans between 1980 and 2022, but generally amounts to available data of around 20–25 years. In Figure S10, we show the temperature anomalies at all five mooring locations from TAO/TRITON and for the corresponding horizontal grid point

in the hindcast simulation. Despite some gaps in the observational data, a comparison between model and observations is feasible. In the observations and the model simulation, the strongest temperature anomalies during the 1997–1998 El Niño are generally found in the subsurface within the upper 200 m (Fig. S10). The model is able to reproduce the initial warming anomalies in the subsurface of the equatorial Pacific and shows an upward migration of the warm anomalies that is similar to observations. Similarly, the model reproduces how warm anomalies subside again first in the subsurface and turn to cold anomalies with the onset of the subsequent La Niña. While the model slightly underestimates the anomalous warming, especially towards the eastern mooring locations, this comparison gives us confidence, that the model realistically reproduces subsurface warming events in the equatorial Pacific.

2. Clustering of MHWs

2.1. Clustering methodology

We cluster the MHWs using a k-means clustering algorithm (Pedregosa et al., 2011), based on a set of their characteristics (see Sec. 2.5 of main text.). To eliminate collinearity between the clustering features (Fig. S11), we perform a principal component analysis (PCA). We find only the first three principal components (PCs) to have eigenvalues above one, which together explain 77.5 % of the variance (Fig. S12). Following Kaiser's rule (Kaiser, 1960), we use the standardized first three PCs for the k-means clustering. We find the optimal number of clusters to be 4 (Fig. S13). This number of clusters maximizes the Calinski-Habarasz score (Caliński & Harabasz, 1974), minimizes the Davies-Bouldin score (Davies & Bouldin, 1979) and marks a clear transition in the within-cluster sum of squared distances (i.e., the Elbow method).

2.2. Clustering results

As described in the main text, we identify four different clusters of deep-reaching MHW (dMHW) vertical propagation types: a) *block-like*, b) *deepening*, c) *shoaling*, and d) *multi-surfacing* MHWs (Fig. 11 of main text). Despite the overall common shape for all extremes, the within-cluster variability of dMHWs is still substantial. Figure S14 shows 4 randomly selected MHW examples for each cluster. Figure S15 shows the clustering results for the MHW characteristics that were selected to be used in the principal component analysis prior to clustering. Both figures highlight the high variability of MHW characteristics within clusters. For comparison, Figure S15 shows the MHW characteristics for the ML-confined MHWs (sMHWs), which were not considered for the clustering.

Fig. S16 shows the clustering results, but for the principal components on which the the clustering was finally performed.

Lastly, Fig. S17 shows the clustering results for secondary MHW characteristics, that is characteristics that are not directly used for clustering, such as the simple column duration, surface duration, the mean depth relative to the MLD, the mean depth relative to the surface, and the mean fraction of the MHW being present in the mixed layer (ML). Again for comparison, the ML-confined sMHW properties are also shown in Figure S17. We find that the distinction between the different clusters is also reflected in these secondary characteristics.

3. Sensitivity Analyses

3.1. MHW characteristics for different MHW detection cases

As outlined in Section 2.6 of the main text, we test for the sensitivity of the detection and characteristics of MHWs regarding seven different methodological choices (see Fig. S18 for an overview). In each sensitivity case we solely alter one methodological choice, while keeping all other choices as in the reference case, which is used throughout the main manuscript.

Figure S19 shows mapped (composite) MHW characteristics for the reference case and all seven sensitivity cases. All characteristics (number of surface-only MHWs per year, number of all MHWs (dMHWs and sMHWs) per year, percentage of MHWs that are dMHWs, mean dMHW duration, mean dMHW depth below the MLD, and the composite maximum of the maximum dMHW intensity) show only little variation between the reference and sensitivity cases. The most striking difference occurs for the sensitivity case D, with fewer, slightly shorter and slightly shallower MHWs than in the reference case (Fig. S19e,m,C,L). However, this is to be expected, as the threshold for the extreme temperature detection is elevated to the 95^{th} percentile in this sensitivity case, leading to substantially less ($\sim$50 %) grid cells harboring extreme conditions.

3.2. Sensitivity of Boolean array B to morphological operations

Applying the morphological operations to smooth the Boolean array $\mathbf{B}$, as outlined in Section 2.2.1 of the main text, has the potential to affect the overall number of detected extreme days. Figure S20 therefore compares the number of extreme days in the unsmoothed and smoothed Boolean array $\mathbf{B}$ at different depth levels (surface, 50 m, 250 m and 500 m depth). For the surface, we perform the same comparison also for the obser-

vational SST data set. We find that in the upper ocean, the effect of the morphological operations on the total number of MHW days is relatively weak. Below 250 m depth, the tropical EP is however marked by local increases of up to more than 20 % in the total number of extreme days (Fig. S20). This implies that the unsmoothed extreme signals in the subsurface tropical EP are often interrupted by short non-extreme periods (shorter than five days). The morphological operations fill these gaps. Still, sensitivity case F (Fig. S18) shows only minor differences in the analyzed MHW characteristics between the smoothed and unsmoothed case (Fig. S19).

3.3. Sensitivity of MHW clustering

We test the robustness of the MHW clusters which we obtained from the k-means clustering described in Section 2.5 of the main text. We therefore analyze how sensitive the clusters are with regards to a) an omission of 10–99 % of MHWs and b) the omission of individual MHW characteristics feeding into the principal component anlysis prior to clustering (Fig. S21). For each case we compare the cluster agreement for the labeled MHWs between the standard case and the sensitivity case using Cohen's Kappa coefficient (κ, Cohen (1960)). $\kappa = 1$ indicates perfect agreement, while $\kappa = 0$ suggests agreement based on random labeling.

As the MHW omission is random, we repeat each sensitivity case (10 % to 99 % of MHWs omitted) ten times, to avoid accidental high agreement between the reference and the respective sensitivity case. As we detect in total 1 400 170 MHWs, an omission of 99 % of MHWs still leaves ∼14 000 MHWs for clustering. For all analyzed MHW omission cases we find on average very high agreements between the cluster assignment for the reference and sensitivity cases ($\kappa \geq 0.99$), indicating high robustness of the clustering. The clustering

consistently assigns the MHWs to the same clusters even under the omission of 99 % of the detected MHWs.

The respective elimination of one of the six MHW characteristics that feed into the principal component analysis, has a somewhat bigger impact on the cluster assignment agreement. In most cases, $\kappa > 0.7$. Only for the omission of the "start delay at the surface" and the "early end at the surface" (normalized with the MHW column duration, see Fig. 2), κ drops to around 0.65, indicating the important role of these characteristics in the clustering procedure. Nevertheless, as all sensitivity cases yield $\kappa > 0.6$, we deem the results of the clustering to be also robust to the general choice of MHW characteristics.

4. Linking surface-only MHWs to their associated subsurface structure

Figure S22 shows two-dimensional histograms between surface-only MHW characteristics and the corresponding characteristics diagnosed for the associated water column MHWs. Overall, the surface-only MHW characteristics are only moderately correlated with the corresponding characteristics of the MHWs in the water column. For instance, the surface-only MHW duration is correlated with a Pearson (Spearman) correlation of 0.46 (0.63) with the associated MHW column duration (Fig. S22a). Correlations for the severity are similar (Fig. S22c) and somewhat lower for the maximum intensity with $r_{\text{Pearson}} = 0.36$ ($r_{\text{Spearman}} = 0.55$, Fig. S22b). Hence, despite these weak to moderate correlations, longer lasting, more intense and more severe surface-only MHWs are generally associated with longer lasting, more intense and more severe MHWs in the water column, respectively. Yet, the two-dimensional histograms show that the longest lasting MHWs in the water column with durations of more than 1000 days are associated with relatively short-lived surface-only MHWs (less than 100 days, Fig. S22a). Similarly,

the most intense and severe MHWs in the water column are associated with relatively moderate surface-only MHWs. These results show the challenges in estimating individual subsurface MHW characteristics based on the surface MHW signature.

In the main text (Section 4.4), we explore the possibility of identifying dMHWs based on the properties of their associated surface-only MHWs. We therefore fit logistic regression models between the binary distinction of the associated MHWs into sMHWs and dMHWs and the associated surface-only MHW properties, such as duration, maximum intensity and severity. Figure 10 of the main text shows the model fit and the predictive capacities of the statistical models fitted for all detected MHWs in the EP. Figure S23 shows the same analysis but for individual subregions of the EP. As such, we separate the EP into eight different subregions by drawing separation lines along 5°S, 5°N, 23°N and 40°N. We distinguish between open ocean regions and coastal regions by using the 700 km distance from the coast isoline. Only in the equatorial Pacific and the subpolar North Pacific (west of 135°W) we do not consider a separate coastal region (Fig. S23). We find that individual regions show higher predictive capacity for the subsurface MHW structure than across the entire EP (compare with Fig. 10 of the main text), indicated by the confusion matrices. While all regions show correct dMHW predictions in around 60–75 %, highest predictive capacity exists in the Equatorial Pacific. There, 77 % (76 %) of MHWs are correctly predicted based on surface-only MHW severity (duration).

References

Boyer, T. P., Garcia, H. E., Locarnini, R. A., Zweng, M. M., Mishonov, A. V., Reagan, J. R., ... Smolyar, I. V. (2018). *World Ocean Atlas 2018.* NOAA National Centers for Environmental Information. Dataset. https://accession.nodc.noaa.gov/NCEI-

WOA18.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. doi: 10.1080/03610927408827101

CMEMS. (2019). Global ocean gridded L4 sea surface heights and derived variables reprocessed (1993-ongoing). *Copernicus Marine Environment Monitoring Service*. doi: 10.48670/moi-00148

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi: 10.1177/001316446002000104

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. doi: 10.1109/TPAMI.1979.4766909

Good, S. A., Martin, M. J., & Rayner, N. A. (2013). EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans*, *118*(12), 6704–6716. doi: 10.1002/2013JC009067

Hobday, A. J., Alexander, L. V., Perkins, S. E., Smale, D. A., Straub, S. C., Oliver, E. C. J., ... Wernberg, T. (2016). A hierarchical approach to defining marine heatwaves. *Progress in Oceanography*, *141*, 227–238. doi: 10.1016/j.pocean.2015.12.014

Holte, J., Talley, L. D., Gilson, J., & Roemmich, D. (2017). An Argo mixed layer climatology and database. *Geophysical Research Letters*, *44*(11), 5618–5626. doi: 10.1002/2017GL073426

Hu, S., Li, S., Zhang, Y., Guan, C., Du, Y., Feng, M., ... Hu, D. (2021). Observed strong

subsurface marine heatwaves in the tropical western Pacific Ocean. *Environmental Research Letters*, *16*(10), 104024. doi: 10.1088/1748-9326/ac26f2

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. doi: 10.1177/001316446002000116

McPhaden, M. J., Busalacchi, A. J., & Anderson, D. L. T. (2010). A TOGA Retrospective. *Oceanography*, *23*(3), 86–103. doi: 10.5670/oceanog.2010.26

Oliver, E. C. J., Benthuysen, J. A., Darmaraki, S., Donat, M. G., Hobday, A. J., Holbrook, N. J., ... Sen Gupta, A. (2021). Marine Heatwaves. *Annual Review of Marine Science*, *13*(1), 313–342. doi: 10.1146/annurev-marine-032720-095144

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825–2830. doi: 10.48550/ARXIV.1201.0490

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., & Schlax, M. G. (2007). Daily High-Resolution-Blended Analyses for Sea Surface Temperature. *Journal of Climate*, *20*(22), 5473–5496. doi: 10.1175/2007JCLI1824.1

Scannell, H. A., Johnson, G. C., Thompson, L., Lyman, J. M., & Riser, S. C. (2020). Subsurface Evolution and Persistence of Marine Heatwaves in the Northeast Pacific. *Geophysical Research Letters*, *47*(23), e2020GL090548. doi: 10.1029/2020GL090548

Sen Gupta, A., Thomsen, M., Benthuysen, J. A., Hobday, A. J., Oliver, E., Alexander, L. V., ... Smale, D. A. (2020). Drivers and impacts of the most extreme marine heatwave events. *Scientific Reports*, *10*, 19359. doi: 10.1038/s41598-020-75445-3

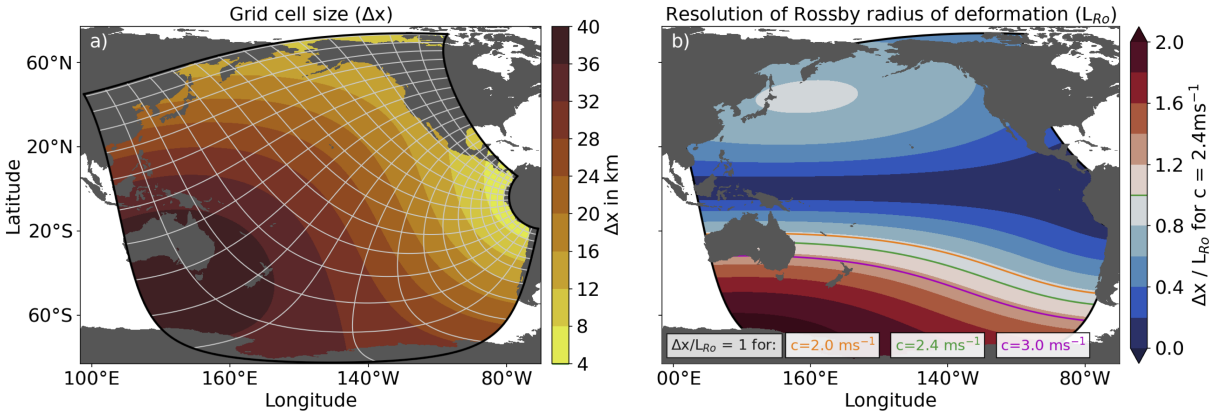

Figure S1. Telescopic grid of the humpac15 Pacific basin setup. Panel a shows the model grid dimension (outlined by black lines) and indicates the grid structure by representing every 50^{th}grid point by the grey lines. The color shows the grid cell size (Δx), with finest resolution off Peru and coarsest resolution off Australia. Panel b shows the ratio of the model grid cell size and the first baroclinic Rossby radius of deformation (L_{Ro}), calculated using a gravity wave speed of $c = 2.4\,\mathrm{m\,s^{-1}}$. Values smaller (larger) than 1 indicate finer (coarser) model resolution than the deformation radius. Orange, green, and magenta lines indicate value of 1 for $c = 2.0$, 2.4, and $3.0\,\mathrm{m\,s^{-1}}$, respectively.

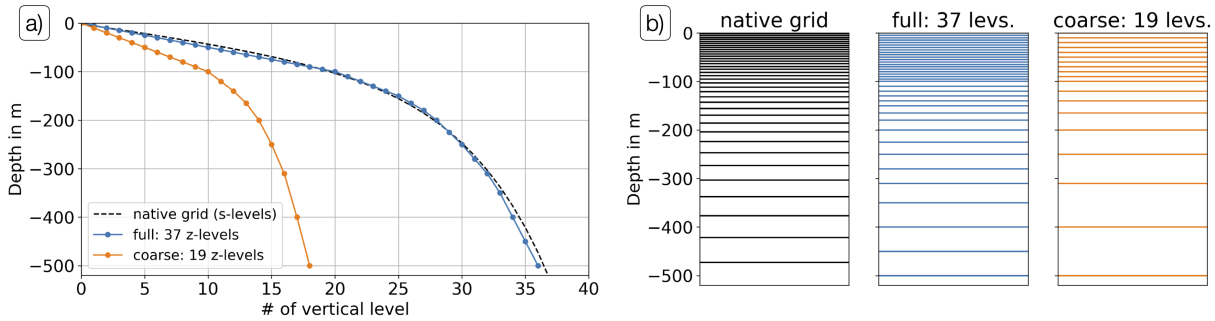

Figure S2. Vertical regridding of model output from terrain following coordinates (s-levels) to fixed z-levels. Panel a shows the vertical levels as a function of depth. The black line shows the model native vertical grid (s-levels) at a location of maximum water depth, i.e. at a maximally stretched vertical grid. The blue and orange lines show the "full" and "coarse" z-level grid, respectively. In the "full" grid, 37 z-levels are chosen to closely follow the maximally stretched s-level depths. The "coarse" z-level grid takes only every second z-level of the "full" grid, leading to 19 z-levels. The different vertical grids are visualized in panel b.

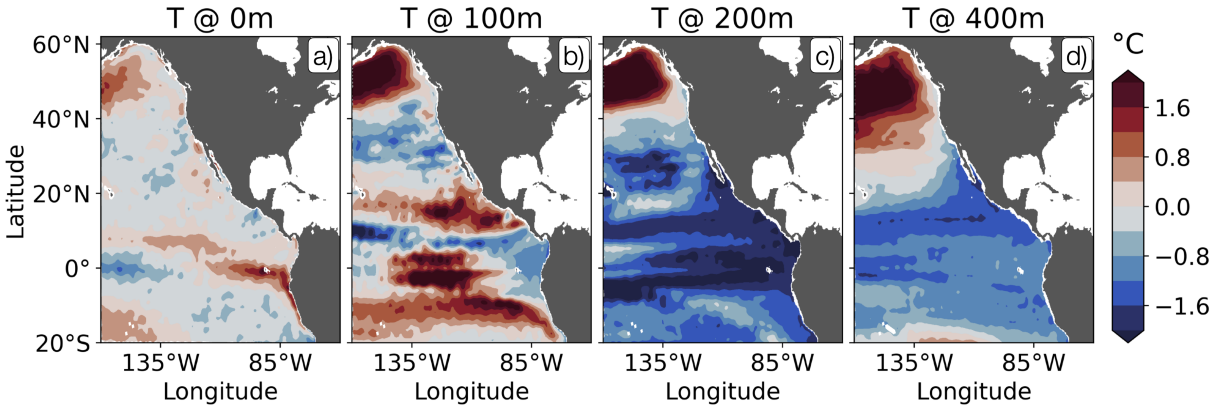

Figure S3. Mean temperature biases of the ROMS model hindcast (1979-2019) at a) the surface, b) 100 m, c) 200 m and d) 400 m depth. Biases are calculated as ROMS minus World Ocean Atlas 2018 (Boyer et al., 2018).

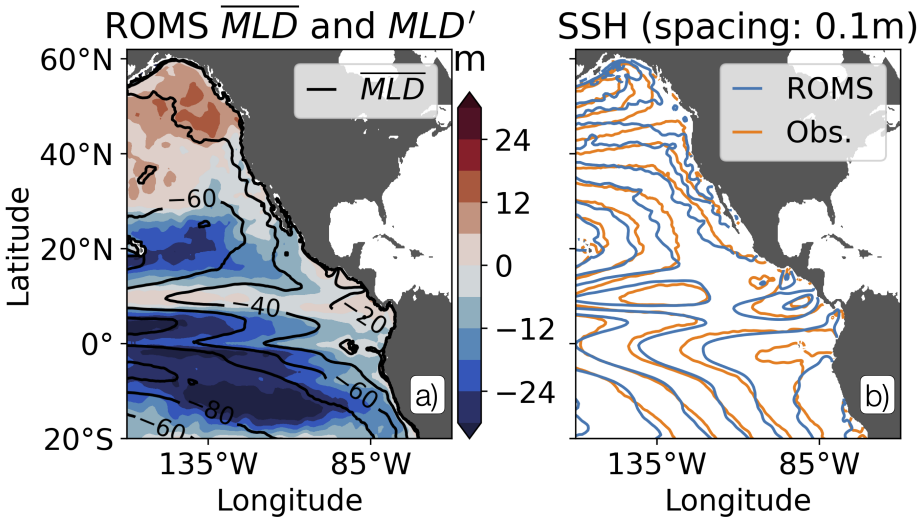

Figure S4. Model evaluation of the mixed layer depth (MLD) and sea surface height (SSH). Panel a shows the hindcast averaged MLD in black contour lines as well as the hindcast averaged MLD bias (MLD') in color (compared to Holte et al. (2017). Panel b compares the hindcast averaged model SSH to satellite altimetry observations (CMEMS, 2019).

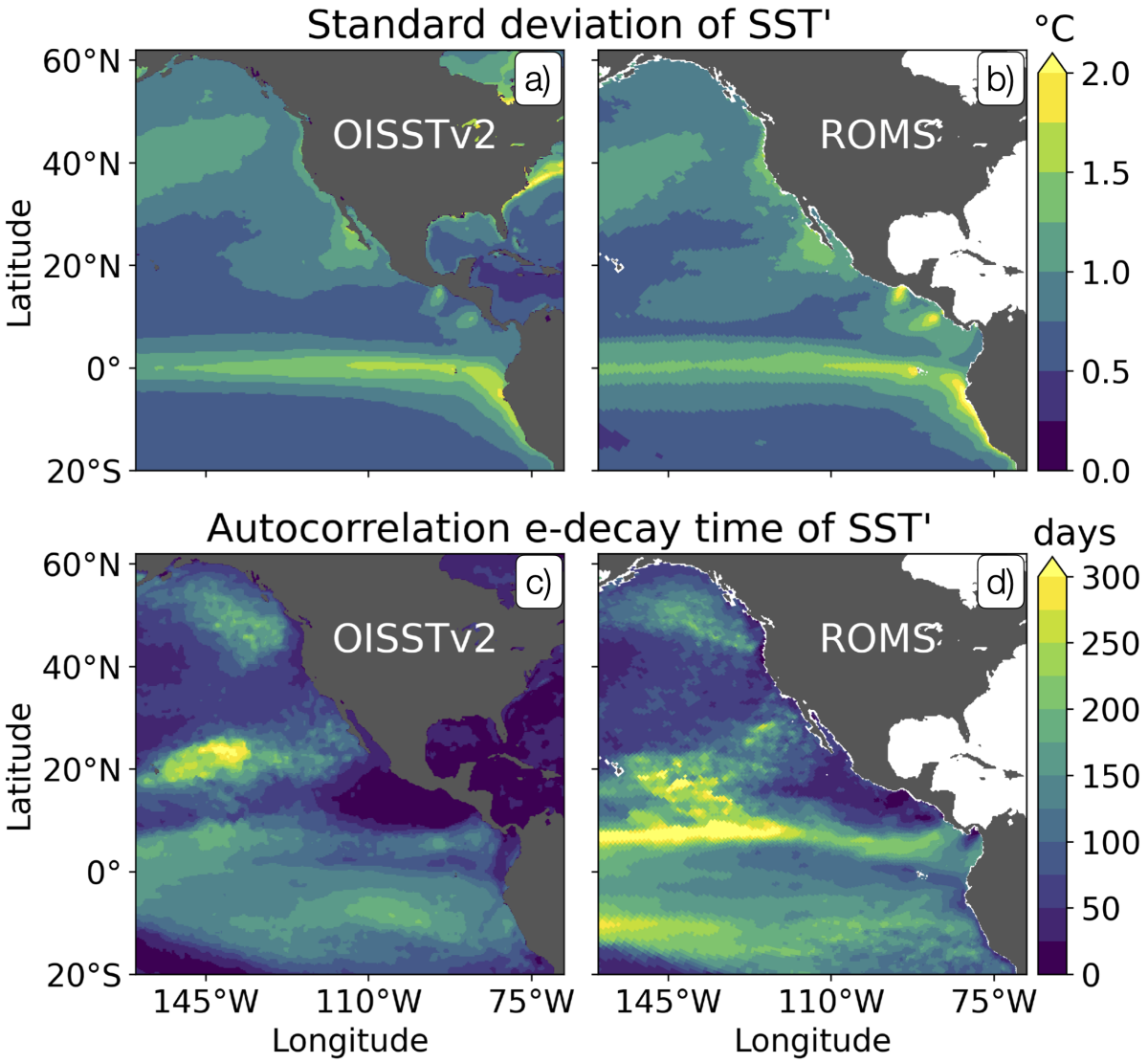

Figure S5. Maps showing the standard deviation (upper) and the auto-correlation e-decay time (lower) of the de-trended SST anomaly (SST') field at the ocean surface. Left panels show the metrics based on observations (OISSTv2 data, Reynolds et al. (2007)), right panels for the ROMS hindcast. Temperature anomalies are calculated relative to the daily climatology of sea surface temperatures calculated over the time period 1982-2011 following Hobday et al. (2016).

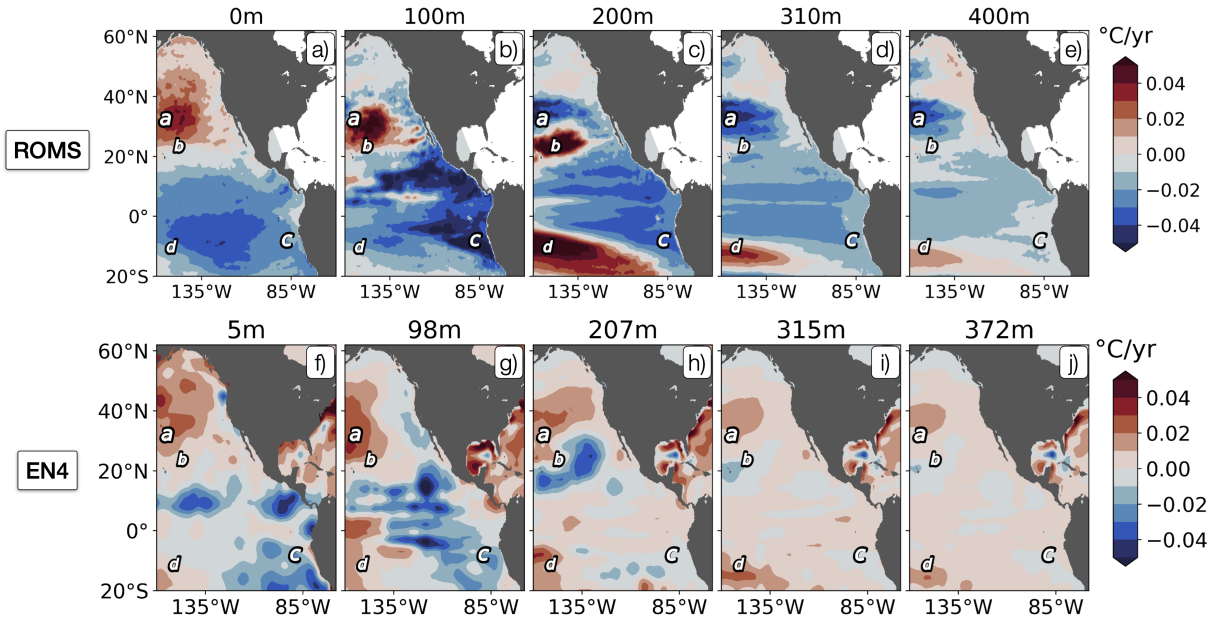

Figure S6. Temperature trends in the ROMS hindcast. Upper row (panels a-e) shows mapped temperature trends calculated over the full hindcast (1979–2019) at different depth levels (surface, 100 m, 200 m, 310 m, 400 m depth). Lower row (panels f-j) shows the same, but derived from the monthly EN4 data set between 1981 and 2019. Shown depths are not identical as in ROMS, but chosen as close as possible. Letters a-d in all panels indicate locations for which temperature time series are shown in Figure S7.

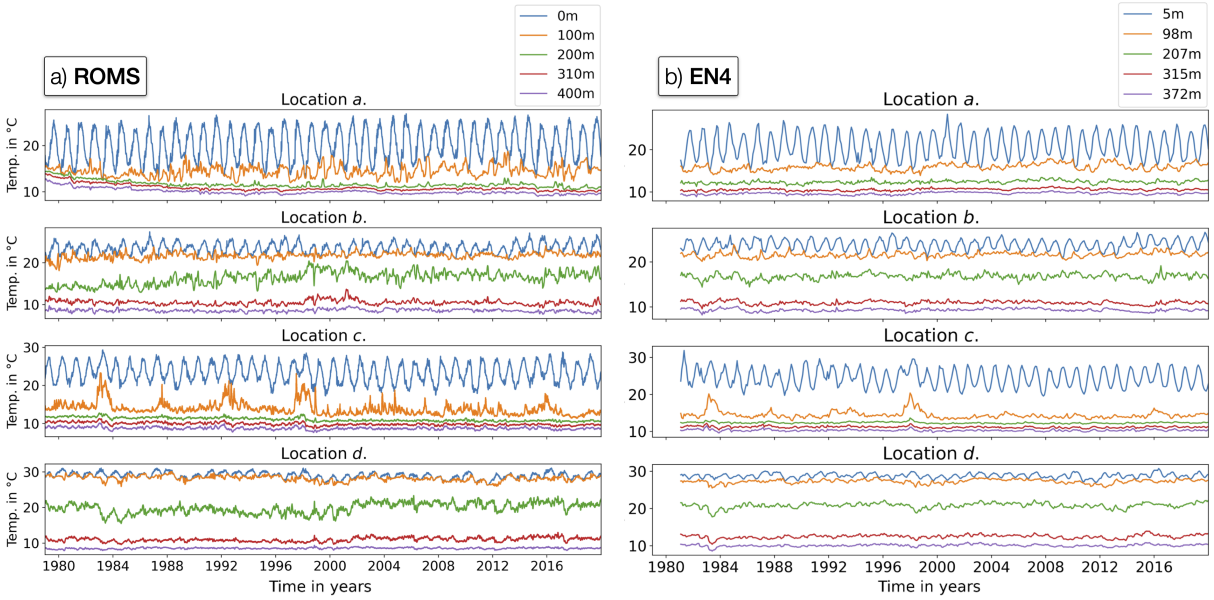

Figure S7. Temperature time series from the daily ROMS hindcast (panel a, 1979-2019) and the monthly EN4 data set (panel b, 1981-2019) at the four locations a-d indicated in Figure S6. For both data sets the time series are shown at for five depths corresponding to the depths shown in Figure S6.

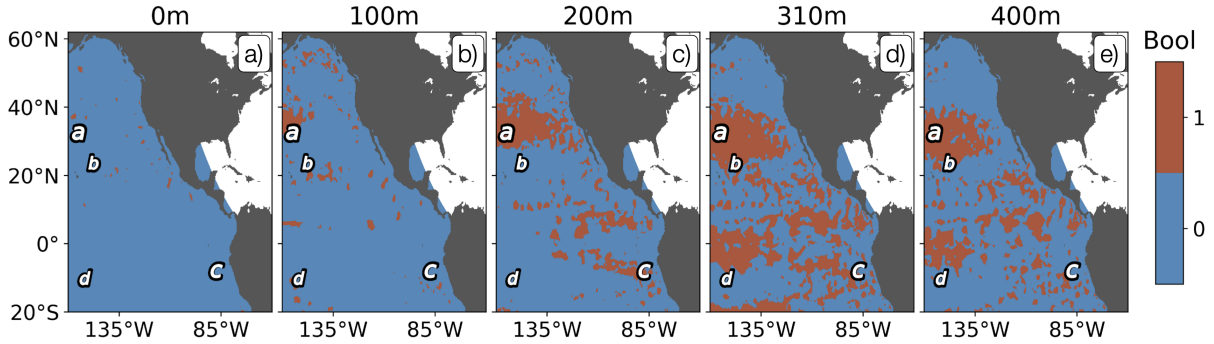

Figure S8. Panels show where extreme temperatures are detected at each depth level on the first day of the hindcast, i.e. Jan 1st, 1979.

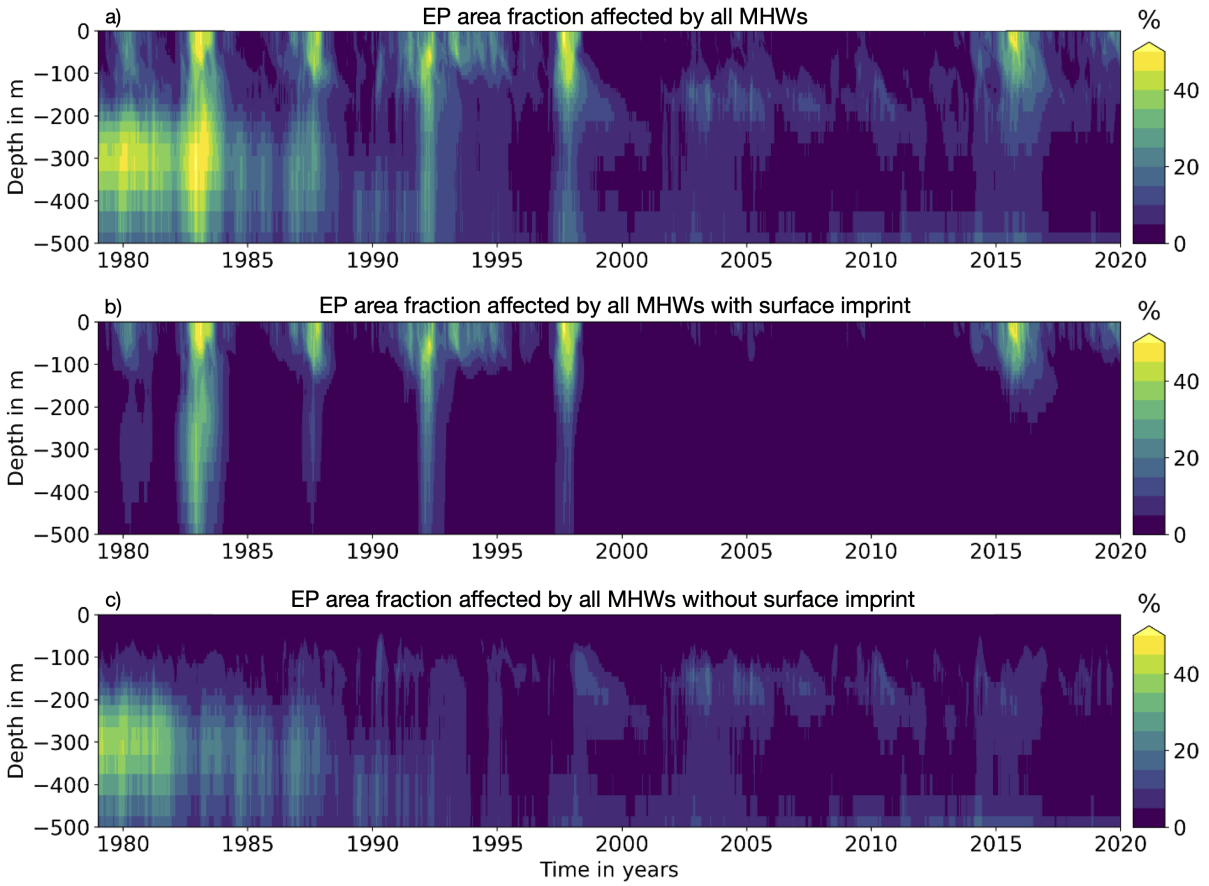

Figure S9. Time series of EP area fraction covered by a) all extreme grid cells, b) extreme grid cells associated with MHWs with a surface imprint, c) extreme grid cells associated with MHWs without a surface imprint (which are thus discarded in this study).

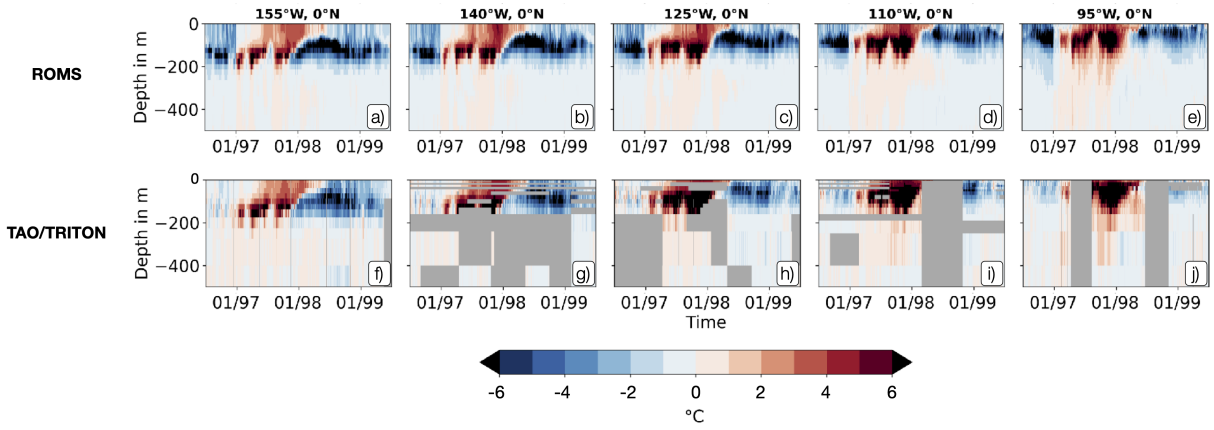

Figure S10. Temperature anomalies relative to the climatology during the 1997 El Niño event as simulated in the ROMS hindcast (top row) and as recorded with the TAO/TRITON array (bottom row) at five different mooring locations along the equator east of 155°W. The panels are arranged by their geographical locations from west to east. Temperature anomalies are calculated relative to the climatology following Hobday et al. (2016). For the hindcast, the climatology is calculated over the 1982–2011 period (as throughout the main study). For the TAO/TRITON data, we use all available data to calculate the climatologies, as the mooring data generally covers only ∼20–25 years.

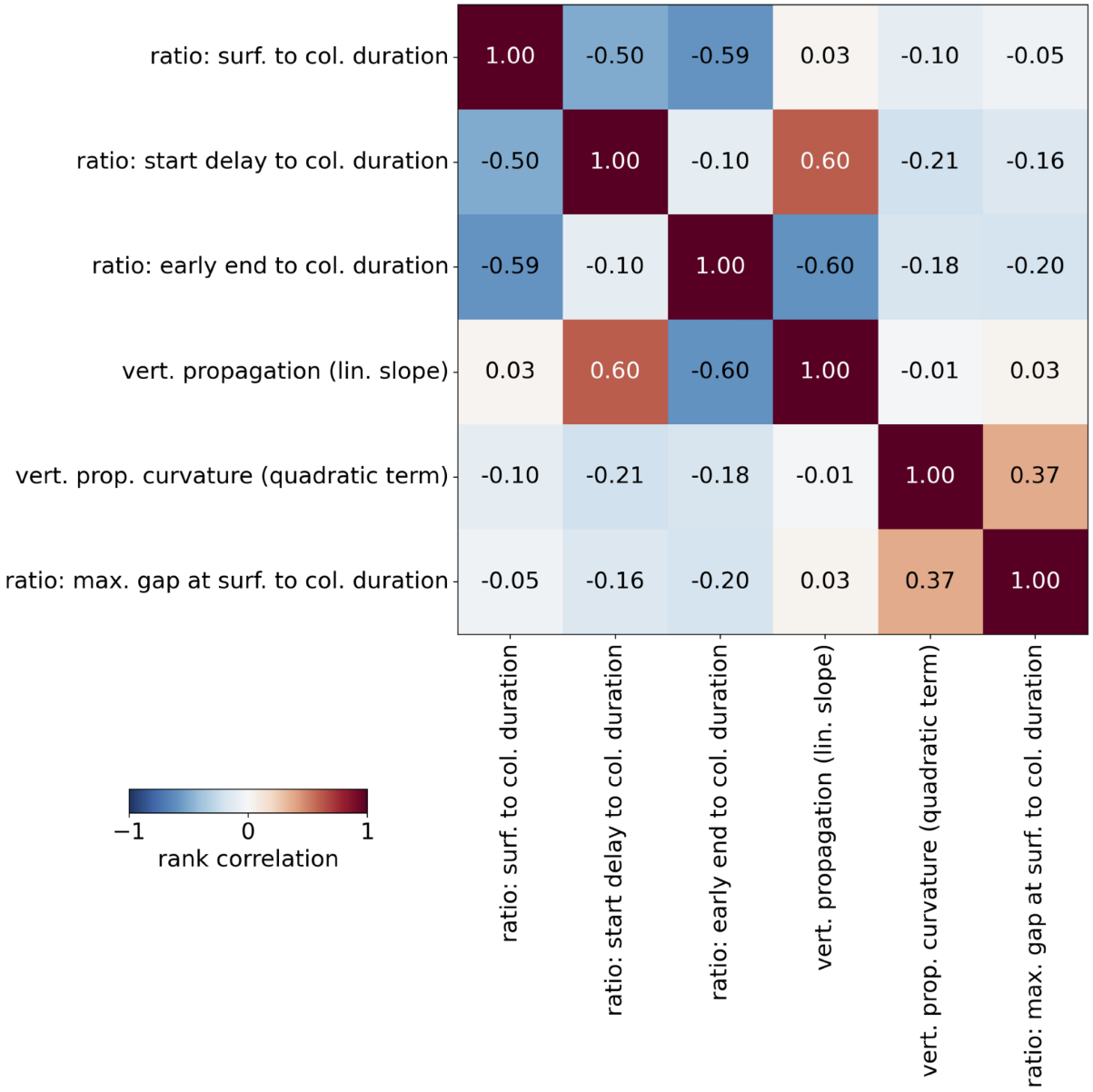

Figure S11. Correlation matrix of primary features, showing rank correlations between all MHW characteristics feeding into the principal component analysis.

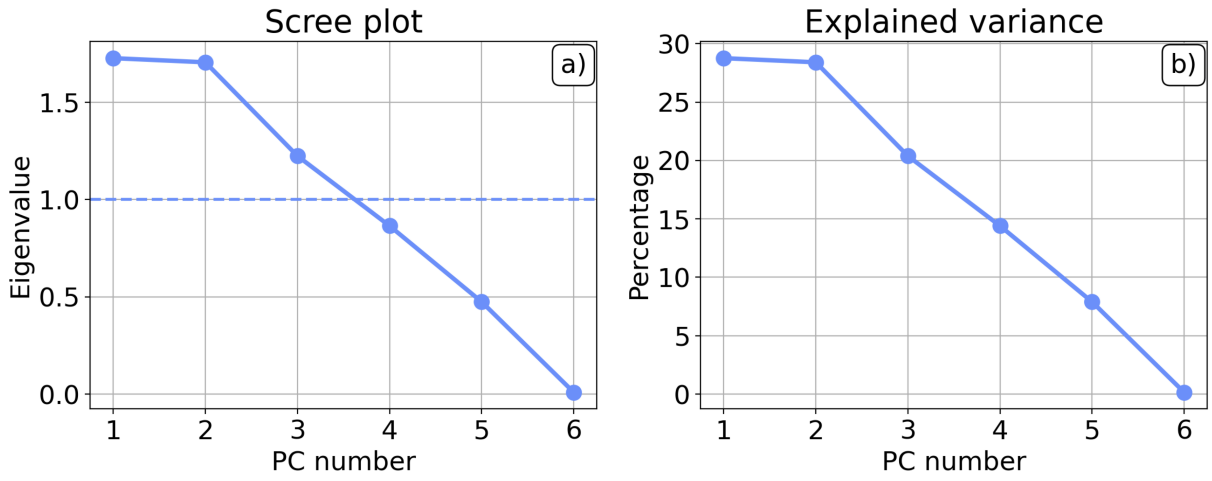

Figure S12. Results of the principal component analysis. Panel a) shows the eigenvalues associated with each individual principal component (Scree plot). Panel b) shows the explained variance by each principal component (PC).

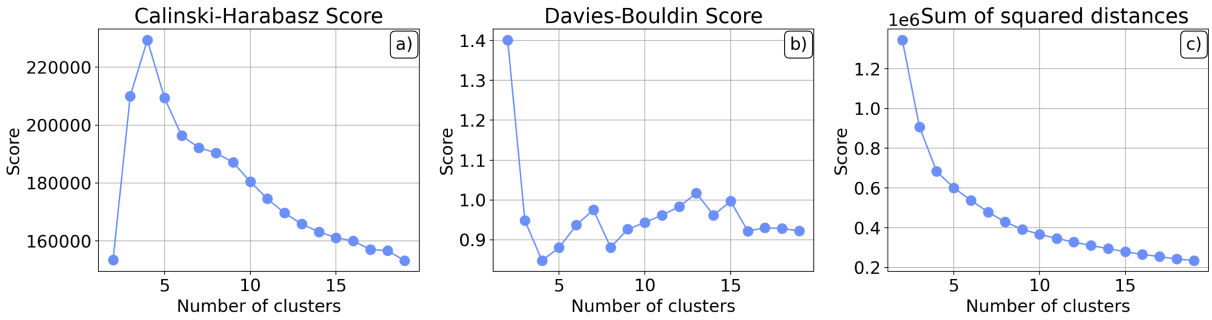

Figure S13. Analysis of optimal number of clusters. Clustering is repeatedly performed using 2–19 clusters. For each chosen number of clusters, the Calinski-Harabasz score (panel a), the Davies-Bouldin score (panel b) and the within-cluster sum of squared distances (Elbow method, panel c) is calculated.

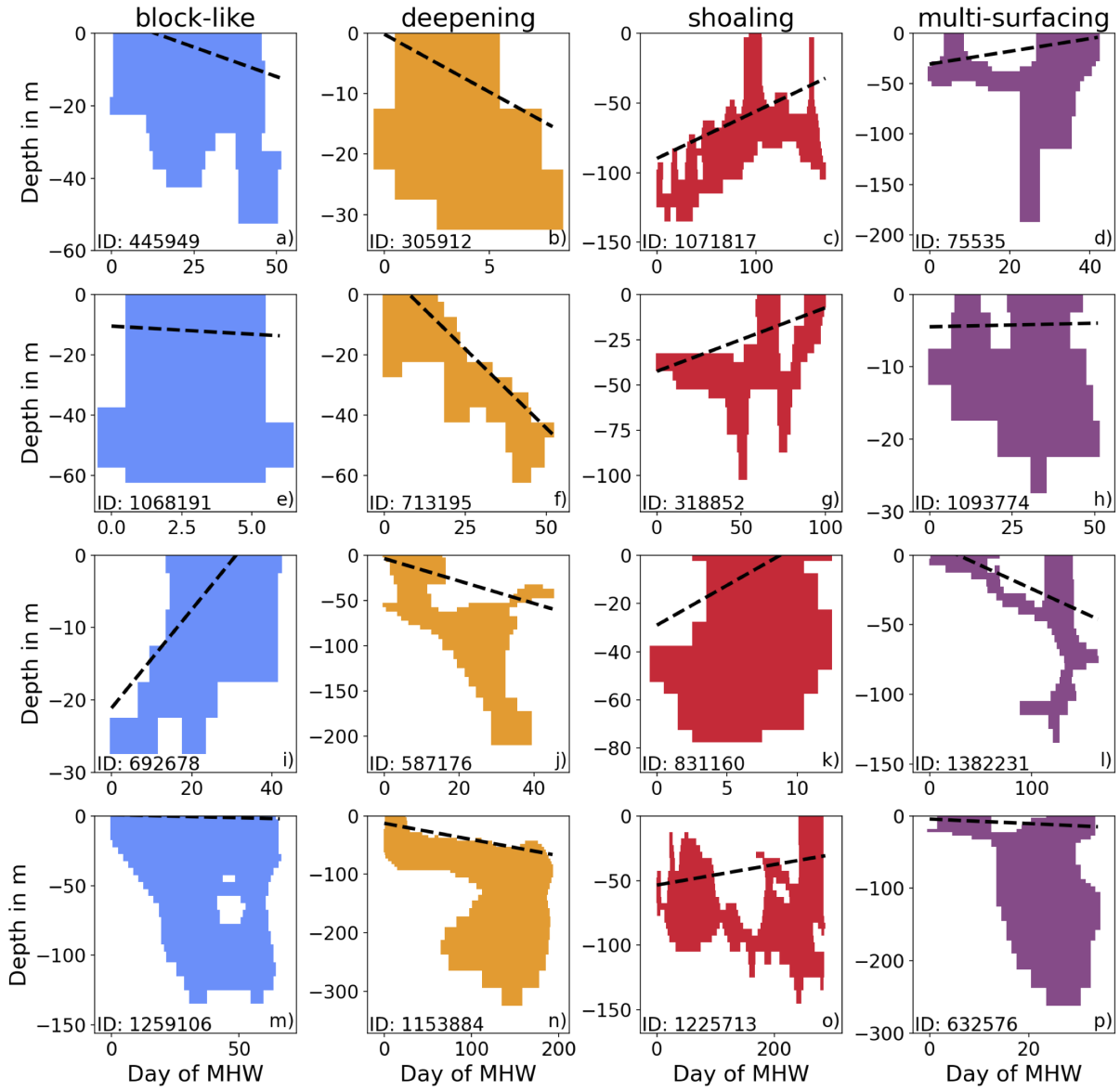

Figure S14. Four examples of the time-depth MHW structures for each cluster, i.e., *block-like* (first column), *deepening* (second column), *shoaling* (third column), and *multi-surfacing* (fourth column) MHWs. Colored areas indicate extreme grid cells. Dashed black line shows linear fit to the temporal evolution of the upper MHW boundary.

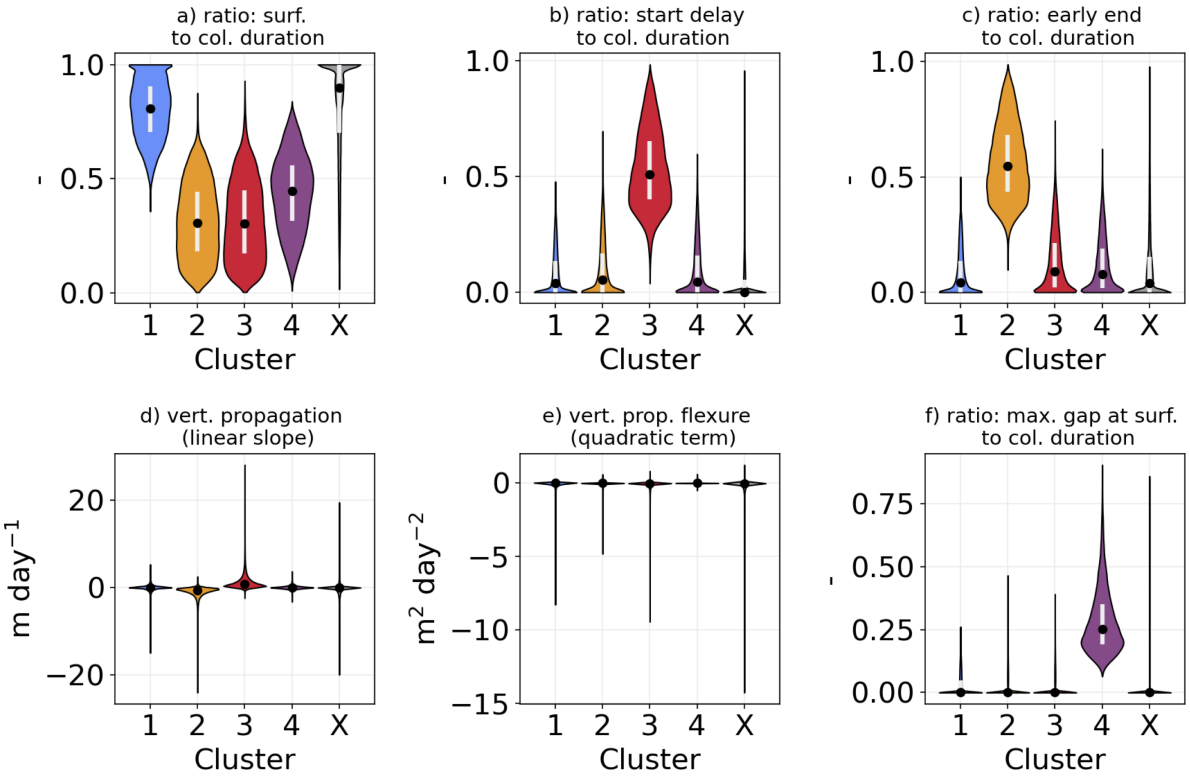

Figure S15. Cluster results. Violinplots show the six MHW characteristics that were used in the principal component analysis based on which the four different clusters were identified (colorcoding/cluster numbering as in Figure 11). Black dots and white lines indicate the median and the interquartile range of the distribution, respectively. For comparison, the corresponding distribution and statistics for the mixed layer-confined sMHWs is shown (denoted by X), even though they are not considered in the clustering.

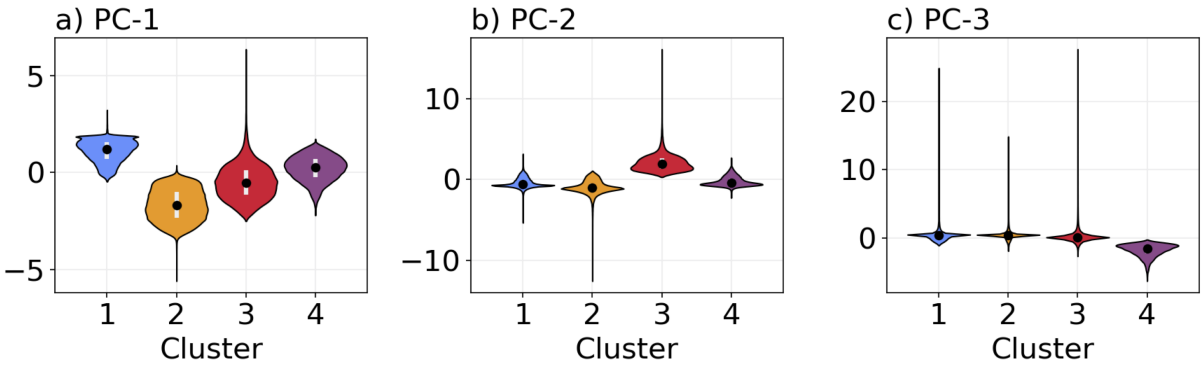

Figure S16. Cluster results. Violinplots show the three clustered PCs (colorcoding/cluster numbering as in Figure 11). Black dots and white lines indicate the median and the interquartile range of the distribution, respectively.

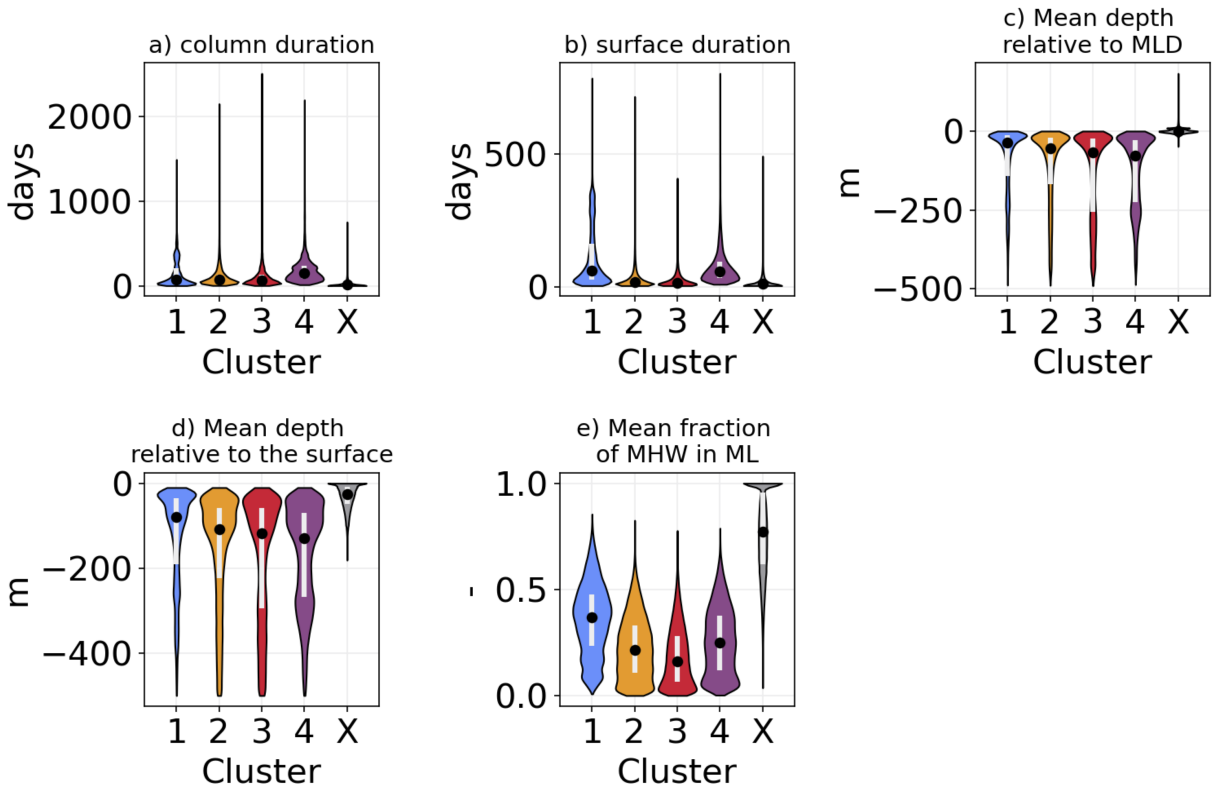

Figure S17. Cluster results. Violinplots show the clustering results for five further MHW characteristics that were not used in the principal component analysis feeding into the clustering (colorcoding/cluster numbering as in Figure 11). Black dots and white lines indicate the median and the interquartile range of the distribution, respectively. For comparison, the corresponding distribution and statistics for the mixed layer-confined sMHWs is shown (denoted by X), even though they are not considered in the clustering.

		Regridded model output			Thresh./clim. array		Bool. array	All arrays
	CASE ID	Vertical depth levels	Horizontal downsampling	Downsampling method	Threshold percentile	Threshold period	Opening/ closing	Analysis period
Reference case	0	37-zlevels (full)	3x3	meanpool	90th percentile	1982-2011	5days	1979-2019
Regridding cases	A	19-zlevels (coarse)	3x3	meanpool	90th	1982-2011	5days	1979-2019
	B	37-zlevels	5x5	meanpool	90th	1982-2011	5days	1979-2019
	C	37-zlevels	3x3	maxpool	90th	1982-2011	5days	1979-2019
Threshold cases	D	37-zlevels	3x3	meanpool	95th	1982-2011	5days	1979-2019
	E	37-zlevels	3x3	meanpool	90th	1990-2019	5days	1979-2019
Smoothing case	F	37-zlevels	3x3	meanpool	90th	1982-2011	None	1979-2019
Period case	G	37-zlevels	3x3	meanpool	90th	1982-2011	5days	1982-2019

Figure S18. The different sensitivity cases regarding the MHW detection. The top row shows the reference case used throughout the main manuscript. In each sensitivity case A-G, only one of the methodological choices outlined in 2.6 is altered (highlighted in red).

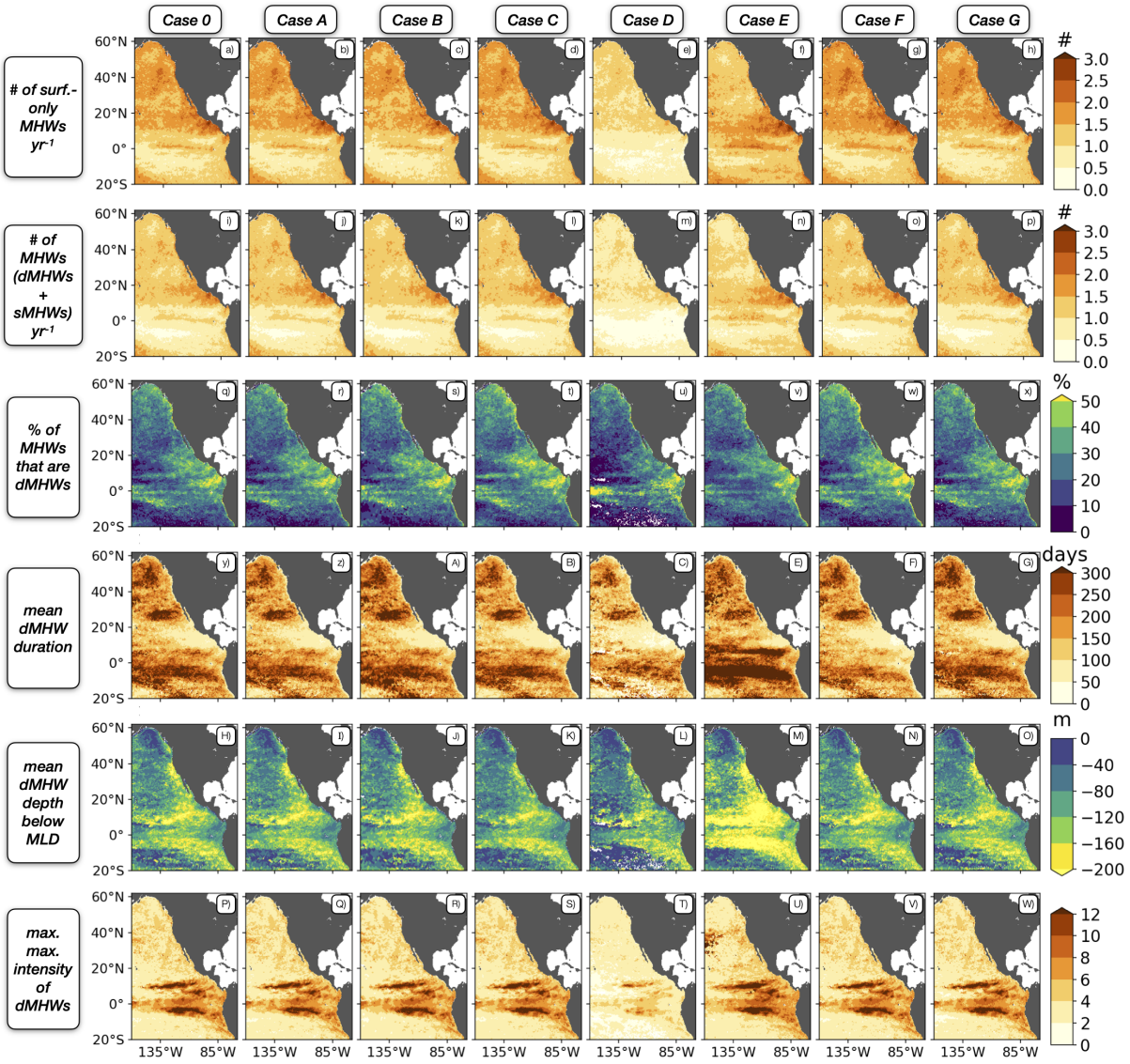

Figure S19. Composite MHW characteristics in different sensitivity cases (see Fig. S18). For comparability, a minimum duration criterion is applied to case F, in which the Boolean array **B** is not filtered using the morphological operations (see Sec. 2.2.1 of main text). It requires that the MHWs have a surface duration of at least five days.

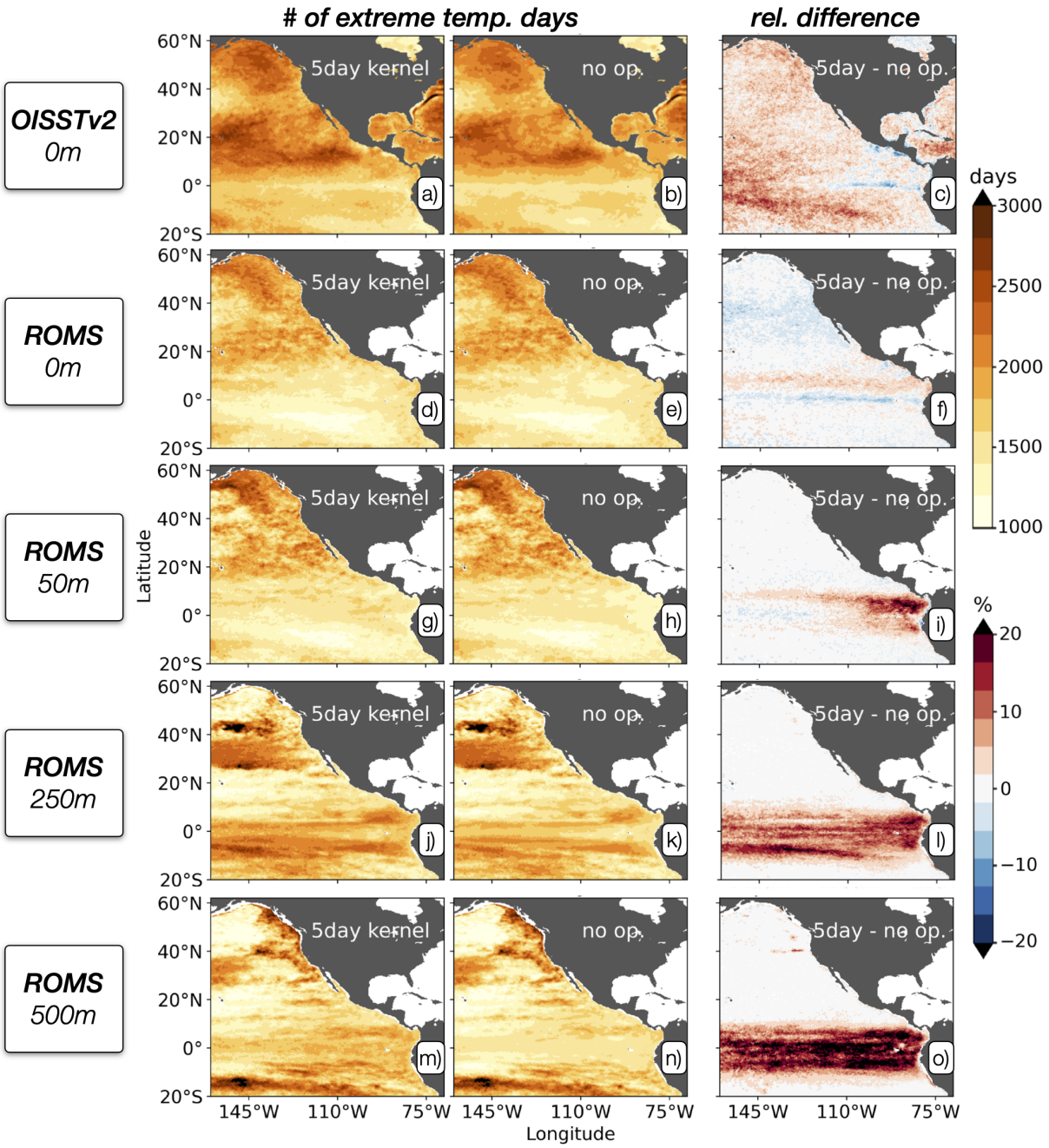

Figure S20. Effect of morphological operations on the number of extreme days in sea surface observations between 1982–2019 (top row, OISSTv2, (Reynolds et al., 2007)) and below in the 1979–2019 ROMS hindcast (surface, 50 m, 250 m and 500 m depth). Left (middle) column shows number of extremes days in smoothed (unsmoothed) Boolean array **B**. Right column shows the relative difference (smoothed minus unsmoothed) in %.

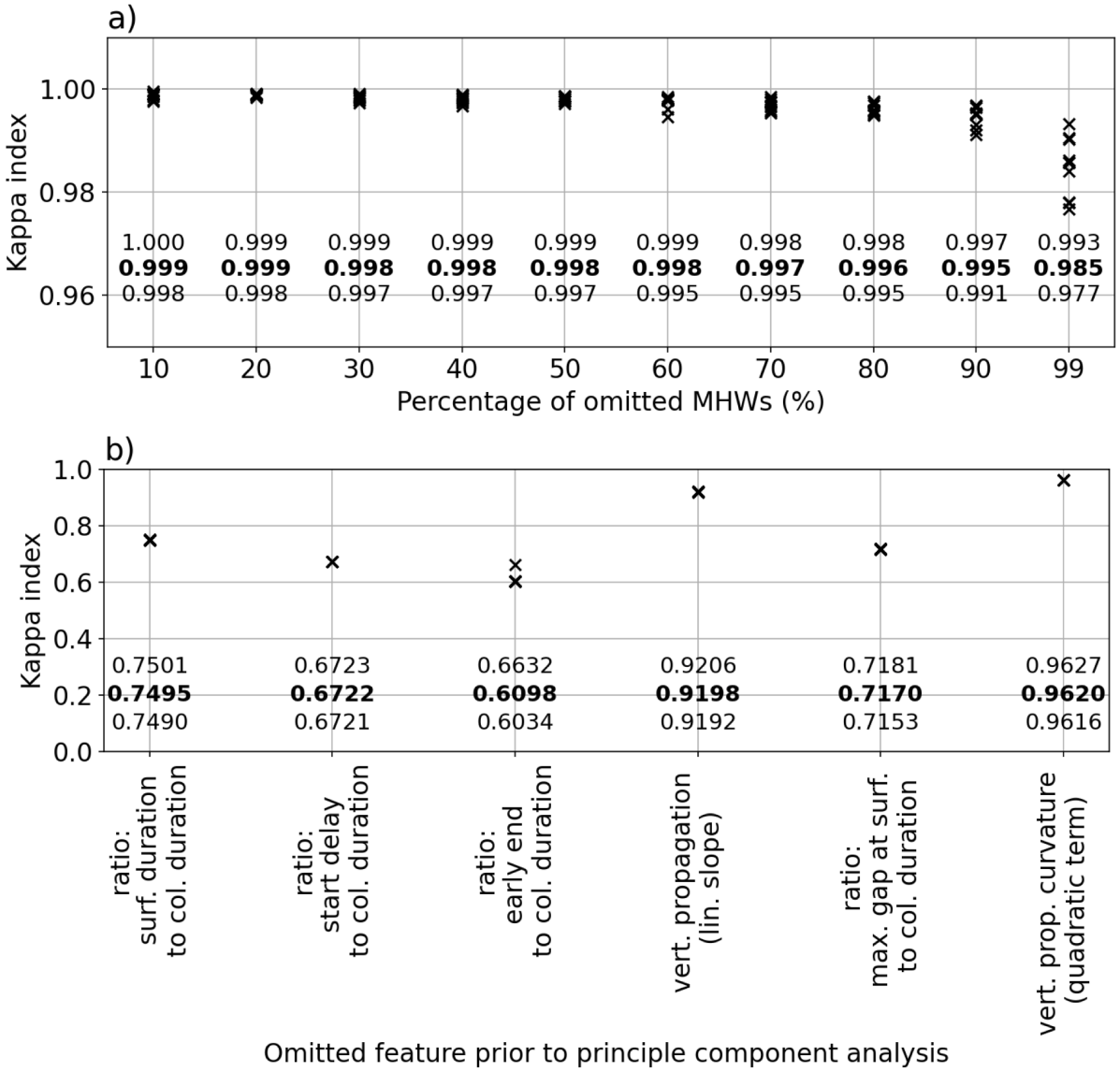

Figure S21. Sensitivity analysis of MHW clustering using Cohen's Kappa coefficient. In the upper panel, we test for the robustness of the clustering with respect to the omission of 10 % to 99 % of omitted MHWs. In the lower panel, we test for the robustness of the clusters with respect to the omission of individual MHW characteristics feeding into the principal component analysis. In both panels we conduct for each case 10 sensitivity clusters. We indicate the mean across all 10 cases by the bold black number and the minimum and maximum value by the numbers below and above, respectively.

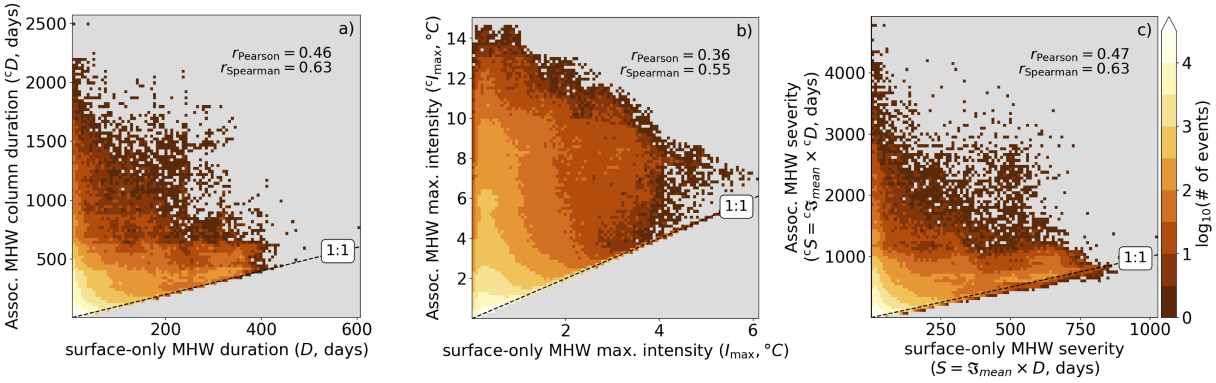

Figure S22. Association between surface-only MHW characteristics and the associated water column MHW characteristics. Panels a, b and c shows each a two-dimensional histogram for the duration (D), maximum intensity (I_{max} in °C) and severity ($S = \mathfrak{I}_{\text{mean}} \times D$ in days) for surface-only MHWs and their associated water column MHWs, respectively. Each panel shows the 1:1 (dashed black) line as well as the Pearson and Spearman correlation (r) between the surface-only and associated water column MHWs.

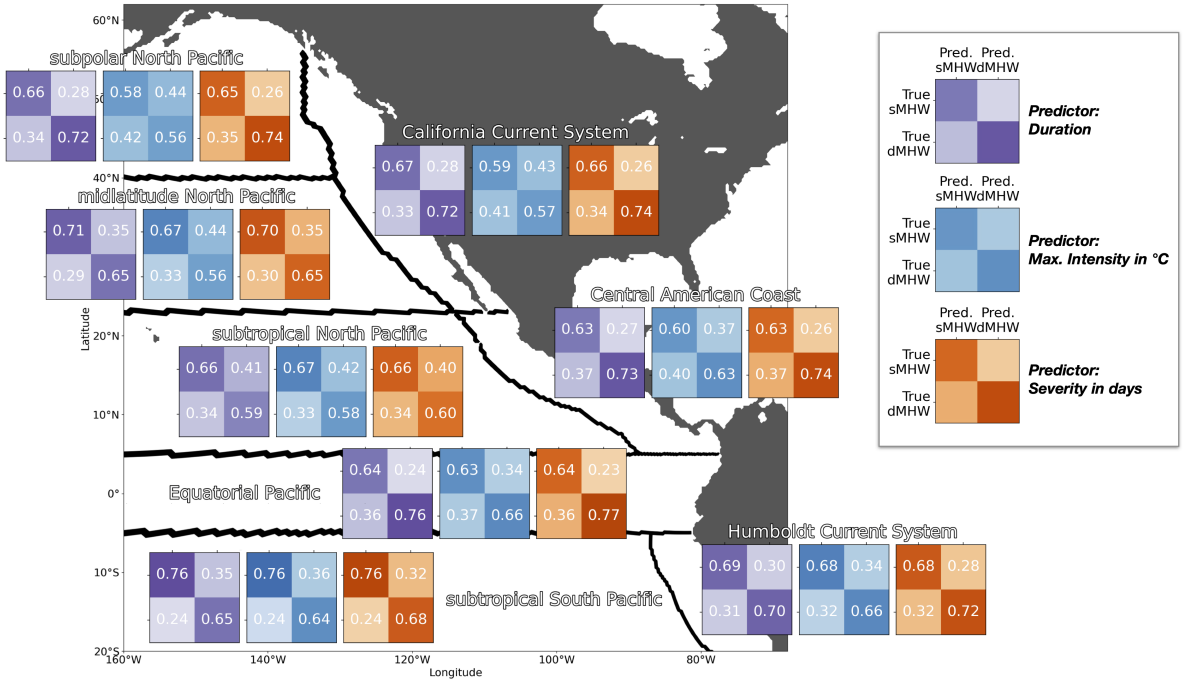

Figure S23. Confusion matrices for regionalized logistic regression model based predictions of MHWs that are either deep-reaching (dMHWs) and ML-confined (sMHWs). Purple, blue and orange matrices show correct/false predictions based on surface only duration, maximum intensity and severity, respectively. Underlying map shows the different regions, demarcated by the black lines. Coastal boxes extend to 700 km offshore. Latitudinal regional boundaries are at 5°S, 5°N, 23°N, and 40°N.