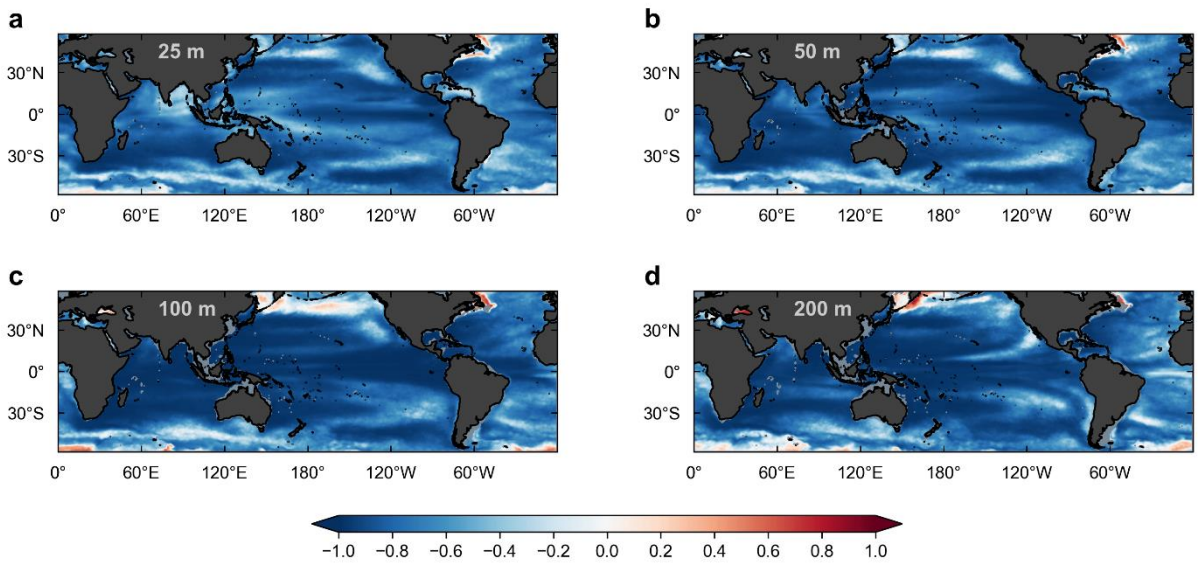
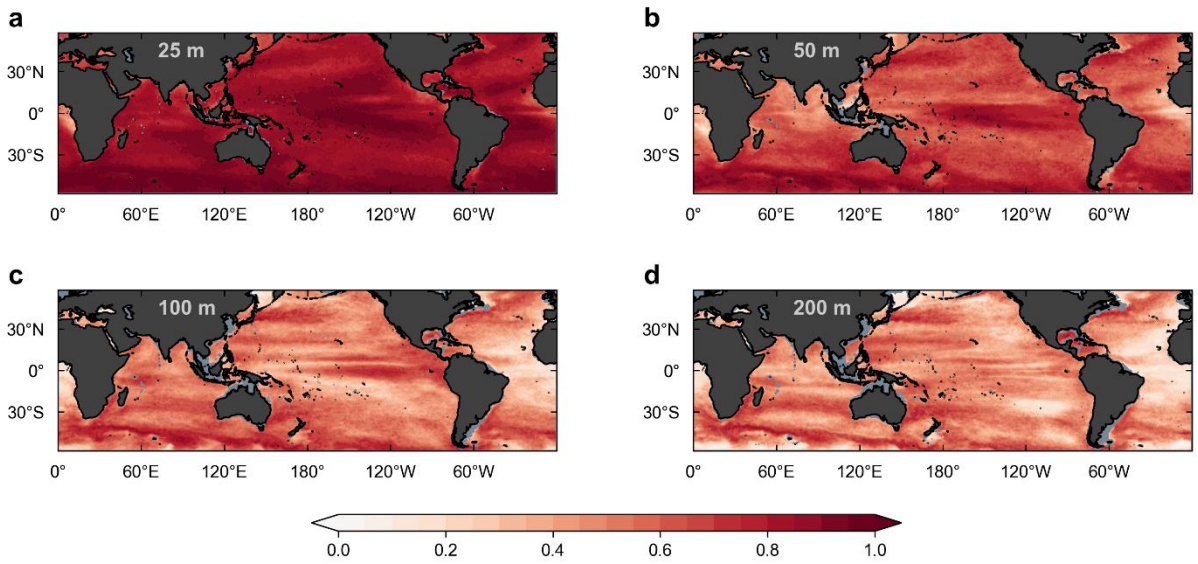
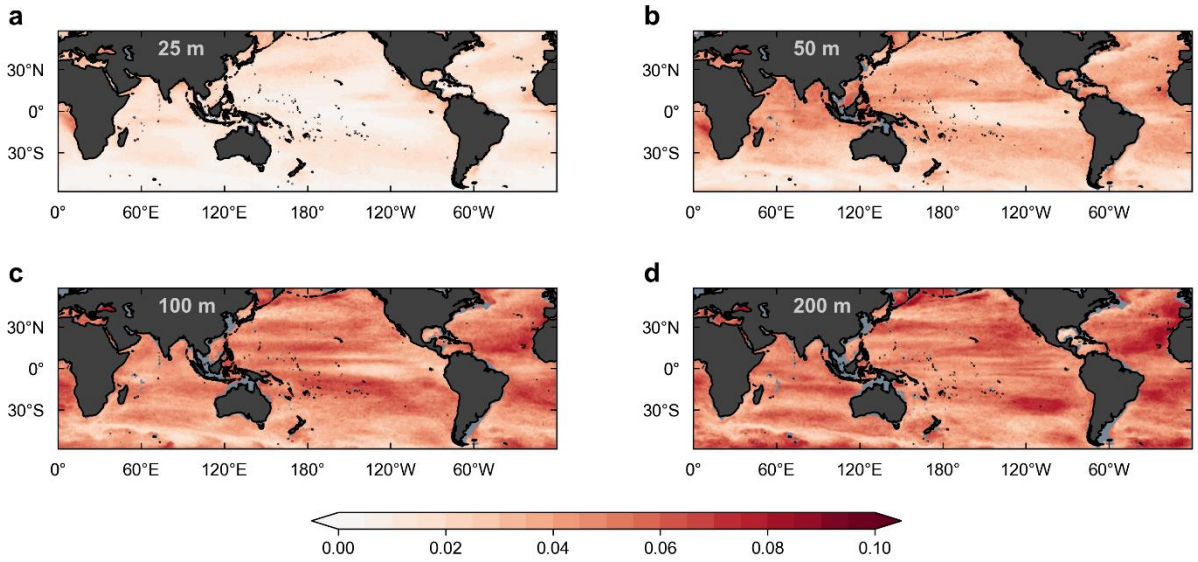


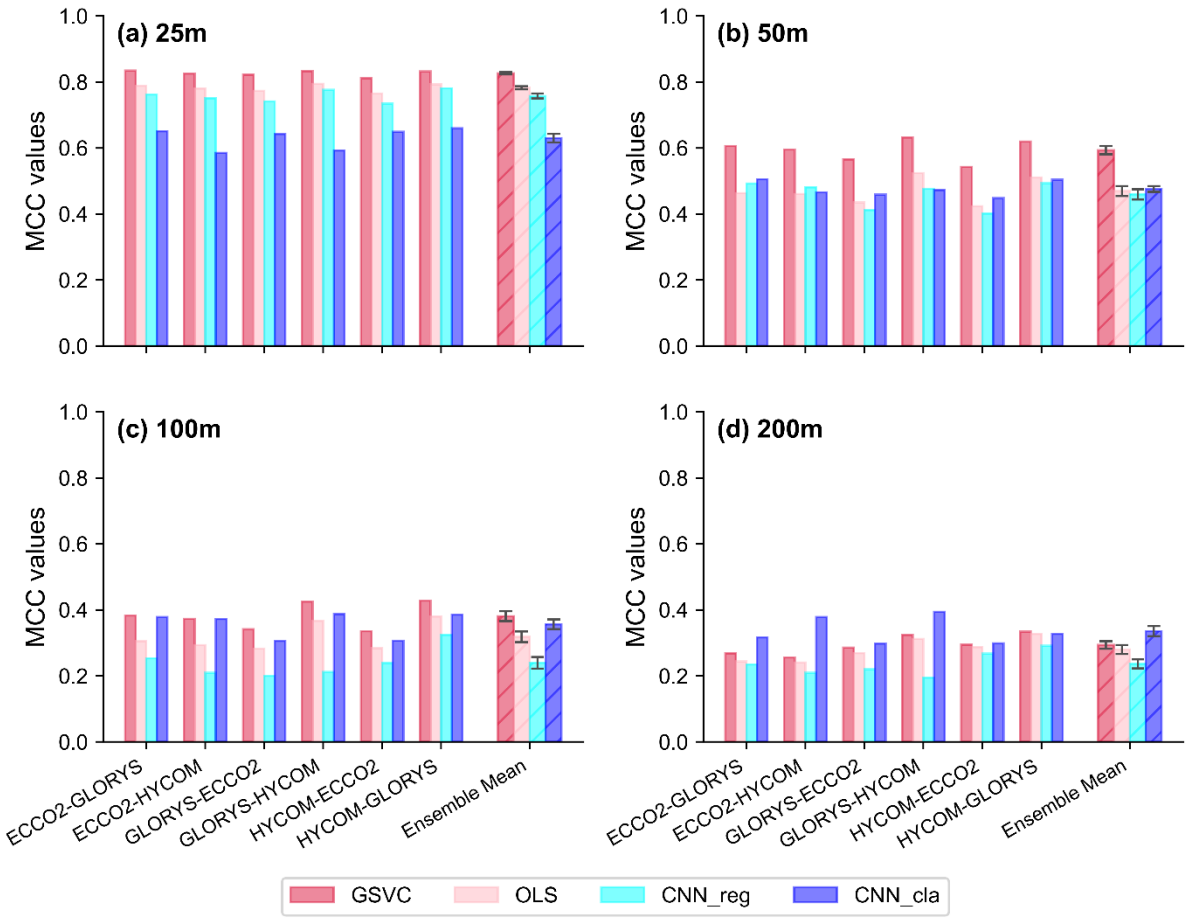
Supplementary Figure 1 | The correlation coefficient between density anomaly and temperature anomaly at 25m (a), 50m (b), 100 (c) and 200 m (d). Here, the correlation coefficient is calculated using GLORYS dataset from 1993 to 2020. The land is marked in dark grey and the continental margins are marked in light grey.



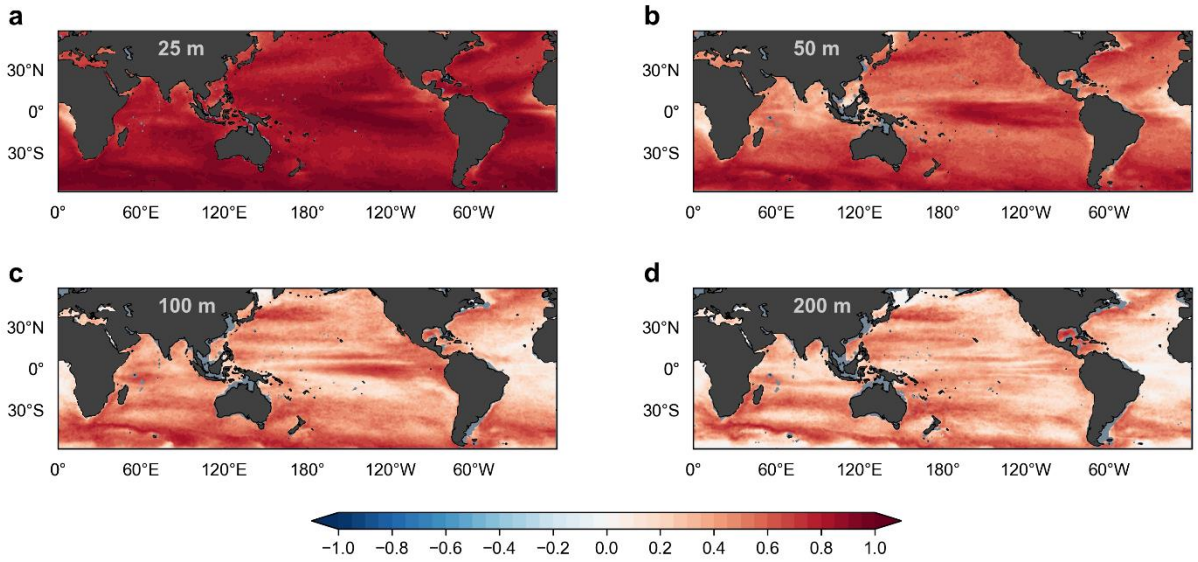
Supplementary Figure 2 | Global distribution of the hit rate for MHW detection at 25 m (a), 50 m (b), 100 m (c) and 200 m (d) averaged over the 6 combinations of training and test datasets. Here the GSVC model is trained for one ocean reanalysis dataset and tested against another ocean reanalysis dataset, leading to 6 combinations of training and test datasets. The land is marked in dark grey and the continental margins are marked in light grey.



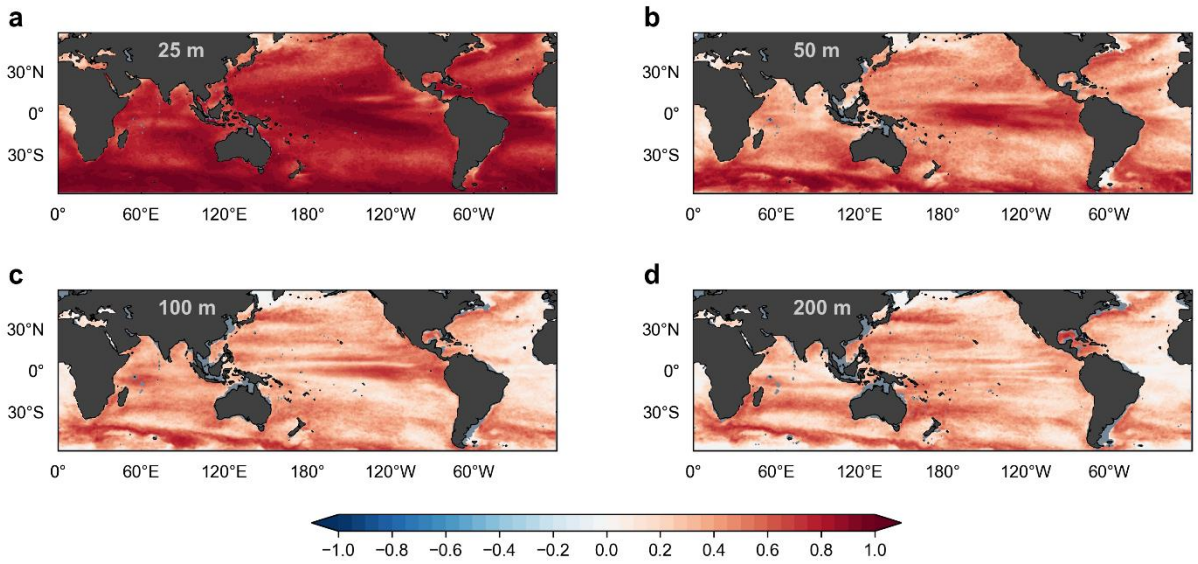
Supplementary Figure 3 | Global distribution of the false alarm rate for MHW detection at 25 m (a), 50 m (b), 100 m (c) and 200 m (d) averaged over the 6 combinations of training and test datasets. Here the GSVC model is trained for one ocean reanalysis dataset and tested against another ocean reanalysis dataset, leading to 6 combinations of training and test datasets. The land is marked in dark grey and the continental margins are marked in light grey.



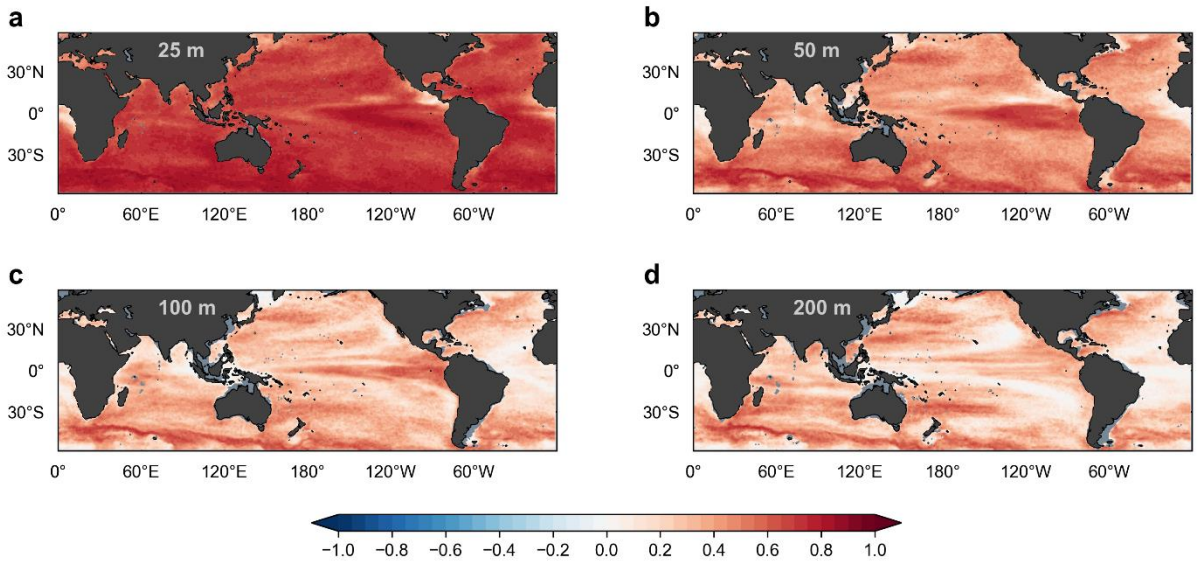
Supplementary Figure 4 | Comparison of performance of the different statistical learning methods in detecting the subsurface MHWs globally between different reanalysis datasets. The global mean (58°S-58°N) test MCC for MHW detection at 25 m (a), 50 m (b), 100 m (c) and 200 m (d) obtained from the GSVC (red), ordinary least square (OLS) regression (pink), and convolutional neural network (CNN) (cyan) models. Here the models are trained and tested on 6 combinations of the different reanalysis datasets. The models are used first to predict subsurface temperature anomaly and then detect MHWs. Direct detection of MHWs based on the CNN model is shown in the dark blue bars. The errorbar is the standard error of the ensemble mean value of 6 combinations.



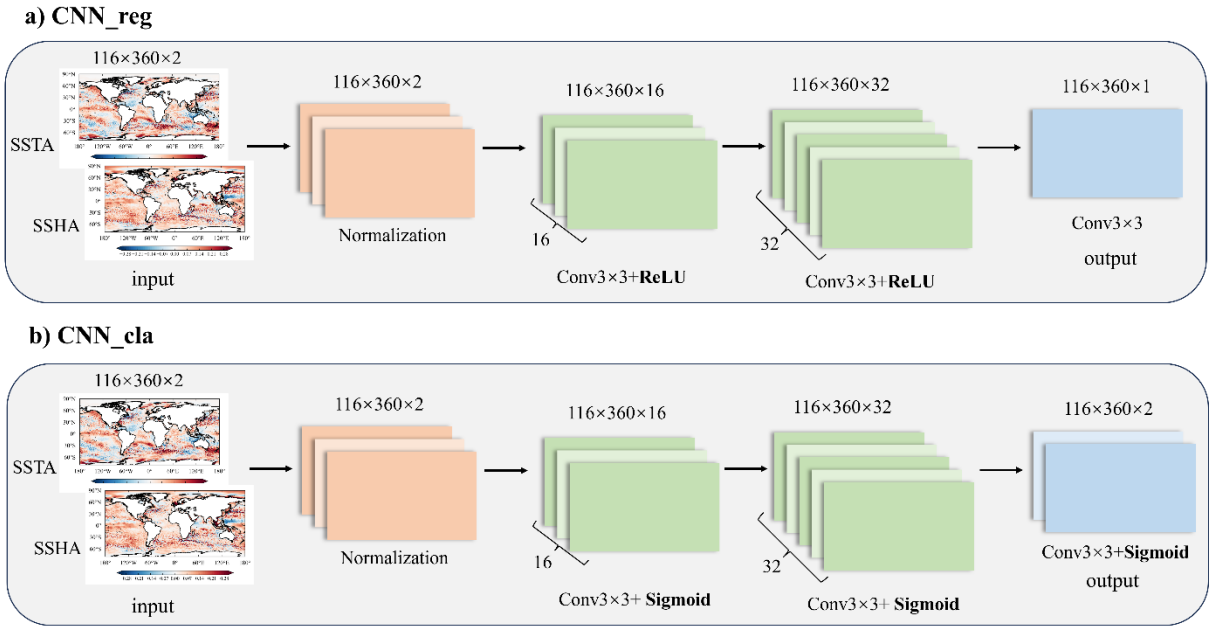
Supplementary Figure 5 | Performance of the GSVC model in detecting the subsurface MHWs globally. Global distribution of the test MCC for MHW detection at 25 m (a), 50 m (b), 100 m (c) and 200 m (d) averaged over three reanalysis datasets. Here the GSVC model is trained and tested on the same reanalysis dataset based on a 5-fold cross validation. The GSVC model is used first to predict subsurface temperature anomaly and then detect MHWs. The land is marked in dark grey and the continental margins are marked in light grey. Area where MCC is not significantly different from zero at the 95% confidence level is filled in white.



Supplementary Figure 6 | Performance of the OLS model in detecting the subsurface MHWs globally. Same as Supplementary Figure 5 but for the OLS model.



Supplementary Figure 7 | Performance of the CNN model in detecting the subsurface MHWs globally. Same as Supplementary Figure 5 but for the CNN model.



Supplementary Figure 8 | A schematic of CNN_reg (a) and CNN_cla (b) models. The orange, green and blue boxes denote the normalization layer, convolutional layer, and fully connected layer, respectively. The dimension of the product at each layer is indicated above the corresponding boxes. The size of filter and the activation function are indicated below the corresponding boxes.

Supplementary Table 1 | Sensitivity experiments on the filter size for CNN_reg (a) and CNN_cla (b). The global mean (58°S-58°N) test MCC for MHW detection at all depths via a 5-fold cross validation on GLORYS reanalysis dataset.

a. CNN_reg

Filter size/° / Depth/m	3 × 3	5 × 5	7 × 7
25	0.6818	0.5934	0.4850
50	0.4725	0.3793	0.4450
100	0.3366	0.3297	0.2994
200	0.2523	0.2475	0.2443

b. CNN_cla

Filter size/° / Depth/m	3 × 3	5 × 5	7 × 7
25	0.6602	0.6486	0.6528
50	0.5069	0.5090	0.4906
100	0.3918	0.3833	0.3836
200	0.3260	0.3217	0.3133

Supplementary Table 2 | Sensitivity experiments on the batch size for CNN_reg (a) and CNN_cla (b). The global mean (58°S-58°N) test MCC for MHW detection at all depths via a 5-fold cross validation on GLORYS reanalysis dataset.

a. CNN_reg

Batch size / Depth/m	256	365	512
25	0.6894	0.6818	0.5646
50	0.4397	0.4725	0.5005
100	0.3197	0.3366	0.3294
200	0.2286	0.2523	0.2318

b. CNN_cla

Batch size / Depth/m	256	365	512
25	0.6597	0.6602	0.6529
50	0.5135	0.5069	0.5054
100	0.3954	0.3918	0.3919
200	0.3312	0.3260	0.3281

Supplementary Table 3 | Sensitivity experiments on the loss function weights for CNN_cla. The global mean (58°S-58°N) test MCC for MHW detection at all depths via a 5-fold cross validation on GLORYS reanalysis dataset.

Weights / Depth/m	(0.1, 0.9)	(0.2,0.8)	(0.3,0.7)
25	0. 6161	0.6597	0.6618
50	0. 4805	0.5135	0.5097
100	0.3647	0.3954	0.3851
200	0.3062	0.3312	0.3203