

# Supplementary Information for A Lightweight Physics-Aware Framework for Multi-Scale Marine Heatwaves Forecasting

Xiu Su,<sup>1,\*</sup> Yuemin Wu,<sup>2,\*</sup> Zhongze Wu,<sup>1,\*</sup> Yitian Long,<sup>3,\*</sup> Yichao Cao,<sup>1</sup>  
Yue Liao,<sup>4</sup> Xi Lin,<sup>5</sup> Jun Long,<sup>1</sup> Shuo Jiang,<sup>6</sup> Shan You,<sup>7</sup> and Chang Xu<sup>2</sup>

<sup>1</sup>*Central South University, Changsha, China*

<sup>2</sup>*University of Sydney, Sydney, Australia*

<sup>3</sup>*Fudan University, Shanghai, China*

<sup>4</sup>*The Chinese University of Hong Kong, Hong Kong, China*

<sup>5</sup>*Shanghai Jiao Tong University, Shanghai, China*

<sup>6</sup>*Tongji University, Shanghai, China*

<sup>7</sup>*SenseTime Research, Beijing, China*

The PDF file includes:

## Supplementary Figures

Supplementary Figure 1. Station specific forecast skill of MARINA at AR.

Supplementary Figure 2. Station specific forecast skill of MARINA at DR.

Supplementary Figure 3. Station specific forecast skill of MARINA at HR.

Supplementary Figure 4. Seasonal anomaly patterns in 2020 across four periods over three reef stations.

Supplementary Figure 5. Seasonal anomaly patterns in 2021 across four periods over three reef stations.

Supplementary Figure 6. Seasonal anomaly patterns in 2022 across four periods over three reef stations.

Supplementary Figure 7. Seasonal anomaly patterns in 2023 across four periods over three reef stations.

Supplementary Figure 8. Additional comparison of forecasting performance across models.

Supplementary Figure 9. Sensitivity of MARINA model performance to training data volume.

Supplementary Figure 10. The full occurrence heatmap between meteorological factors and operators.

Supplementary Figure 11. The detailed prompt for LLM-enhanced neural network design.

Supplementary Figure 12. Multi-resolution forecast skill of MARINA with U/V current for MHW events.

Supplementary Figure 13. Multi-resolution forecast skill of MARINA with MJO/ONI index for MHW events.

Supplementary Figure 14. Multi-resolution forecast skill of MARINA with OHC for MHW events.

Supplementary Figure 15. The detail pipeline of LLM enhanced neural network design.

## Supplementary Tables

Supplementary Table 1. Comparison between the natural variability of the observed SST and the prediction error of MARINA across time resolutions.

Supplementary Table 2. Performance comparison of prediction resolutions under different temporal aggregation strategies.

Supplementary Table 3. Comparative performance of MARINA under limited and full-data training regimes.

Supplementary Table 4. Description of predefined experimental operators.

Supplementary Table 5. Quantitative performance comparison between MARINA and the Diurnal Cycle baseline.

Supplementary Table 6. Publicly accessible high-frequency global data sources compatible with MARINA framework.

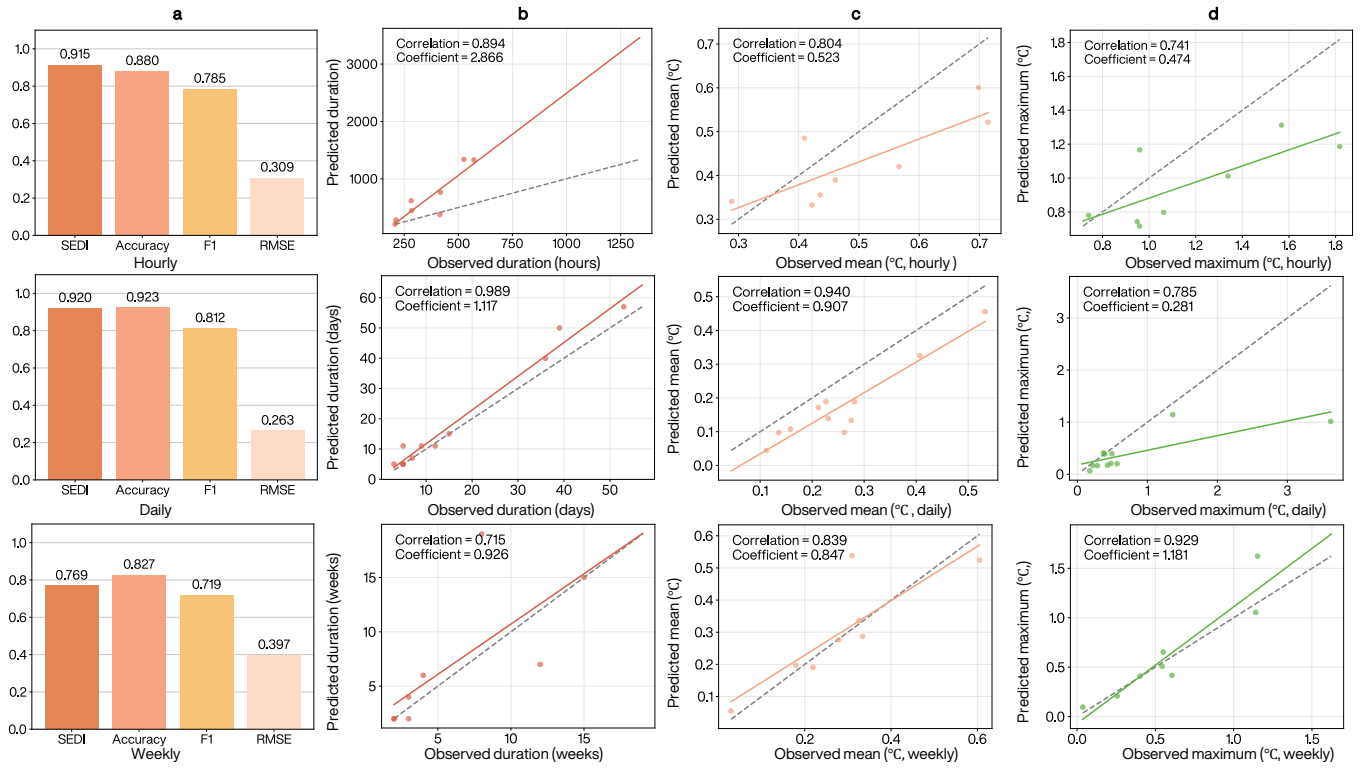
Supplementary Table 7. Summary of data distribution across temporal resolutions and observation sites.

## Supplementary Algorithms

Supplementary Algorithm 1. MT-MHW Data Processing and Labeling Pipeline.

---

\* These authors contributed equally.

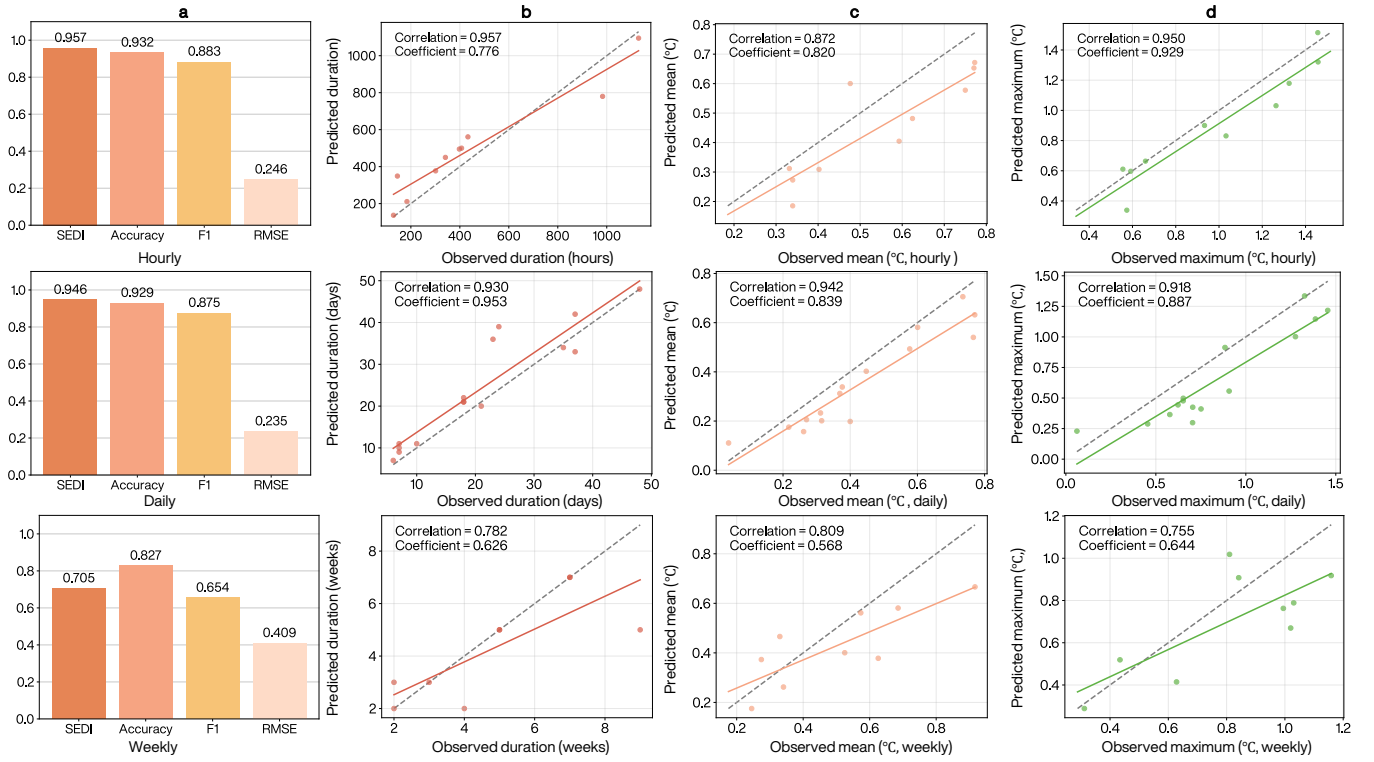


**Supplementary Figure 1:** Station specific forecast skill of MARINA at AR. a, Forecast skill across three temporal resolutions (hourly, daily, weekly) evaluated using SEDI, accuracy, F1 score, and RMSE (top to bottom). b, Predicted versus observed marine heatwave duration for hourly, daily, and weekly events. c, Predicted versus observed mean marine heatwave intensity for the three resolutions. d, Predicted versus observed maximum marine heatwave intensity for the three resolutions. In panels b to d, each point represents one marine heatwave event, the grey dashed line shows the one to one reference line, and the coloured solid line shows the regression fit, with the reported correlation and regression coefficients summarizing the agreement between predictions and observations.

**Supplementary Table 1:** Comparison between the natural variability of the observed (Obs) SST (represented by Standard Deviation, SD) and the prediction error of MARINA (RMSE) across three temporal resolutions. The significantly lower RMSE compared to SD indicates that the model captures physical dynamics rather than the climatological mean.

| Station        | Hourly (°C) |               | Daily (°C) |               | Weekly (°C) |               |
|----------------|-------------|---------------|------------|---------------|-------------|---------------|
|                | Obs. SD     | MARINA RMSE   | Obs. SD    | MARINA RMSE   | Obs. SD     | MARINA RMSE   |
| AR             | 1.6068      | <b>0.3087</b> | 1.6015     | <b>0.2631</b> | 1.5947      | <b>0.3968</b> |
| DR             | 2.0048      | <b>0.2463</b> | 2.0007     | <b>0.2351</b> | 2.4600      | <b>0.4091</b> |
| HR             | 2.0228      | <b>0.1753</b> | 2.1985     | <b>0.1332</b> | 2.1936      | <b>0.3579</b> |
| <b>Average</b> | 1.8058      | <b>0.2434</b> | 1.8011     | <b>0.2105</b> | 2.0273      | <b>0.3879</b> |

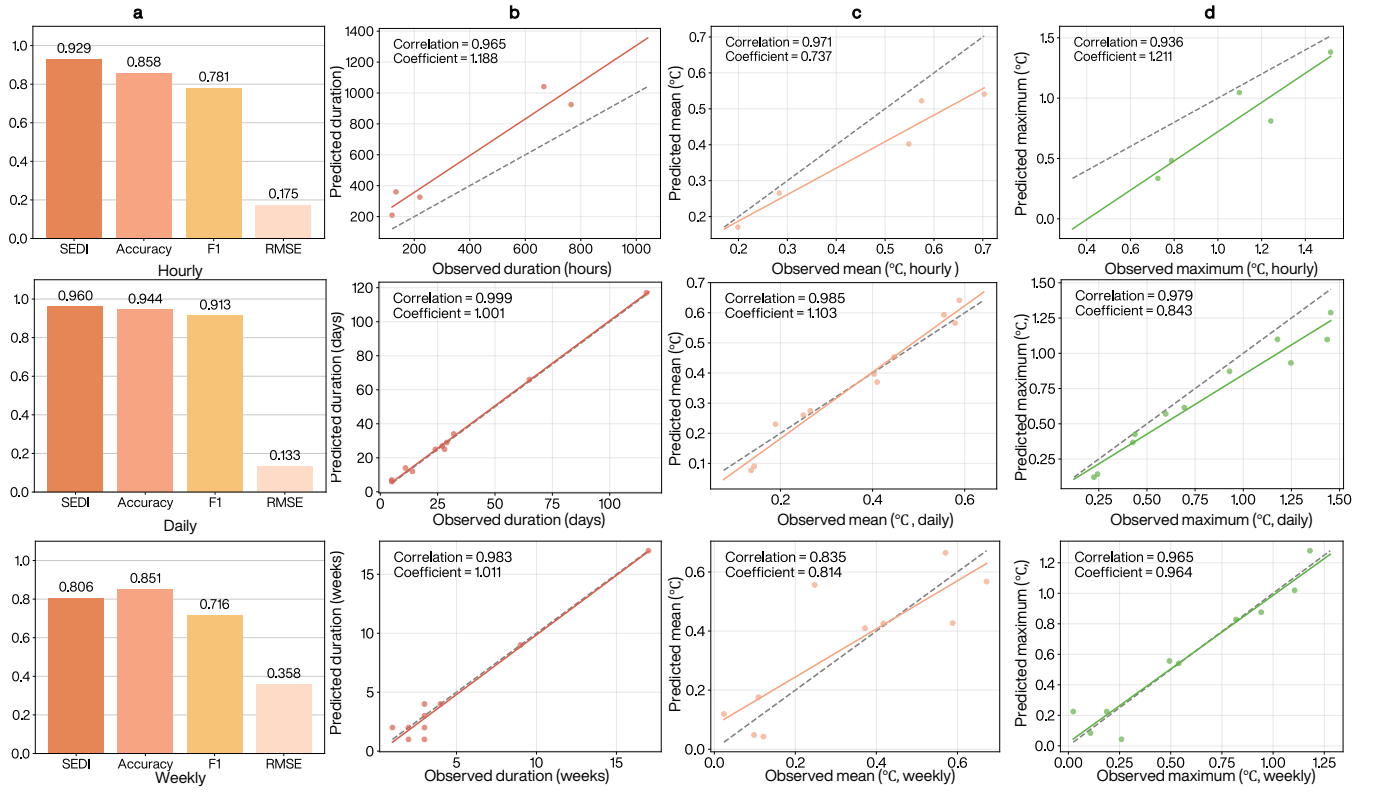




**Supplementary Figure 2:** Station specific forecast skill of MARINA at DR. a, Forecast skill across three temporal resolutions (hourly, daily, weekly) evaluated using SEDI, accuracy, F1 score, and RMSE (top to bottom). b, Predicted versus observed marine heatwave duration for hourly, daily, and weekly events. c, Predicted versus observed mean marine heatwave intensity for the three resolutions. d, Predicted versus observed maximum marine heatwave intensity for the three resolutions. In panels b to d, each point represents one marine heatwave event, the grey dashed line shows the one to one reference line, and the coloured solid line shows the regression fit, with the reported correlation and regression coefficients summarizing the agreement between predictions and observations.

**Supplementary Table 2: Performance comparison of prediction resolutions under different temporal aggregation strategies.** Higher values (↑) indicate better performance for ACC, F1, and SEDI, while lower values (↓) indicate better performance for RMSE, MSE, and MAE. Difference rows (Δ) report the decrease caused by averaging. Our results demonstrate that aggregating fine-resolution forecasts into coarser resolutions leads to a significant decrease in predictive skill compared to direct modeling

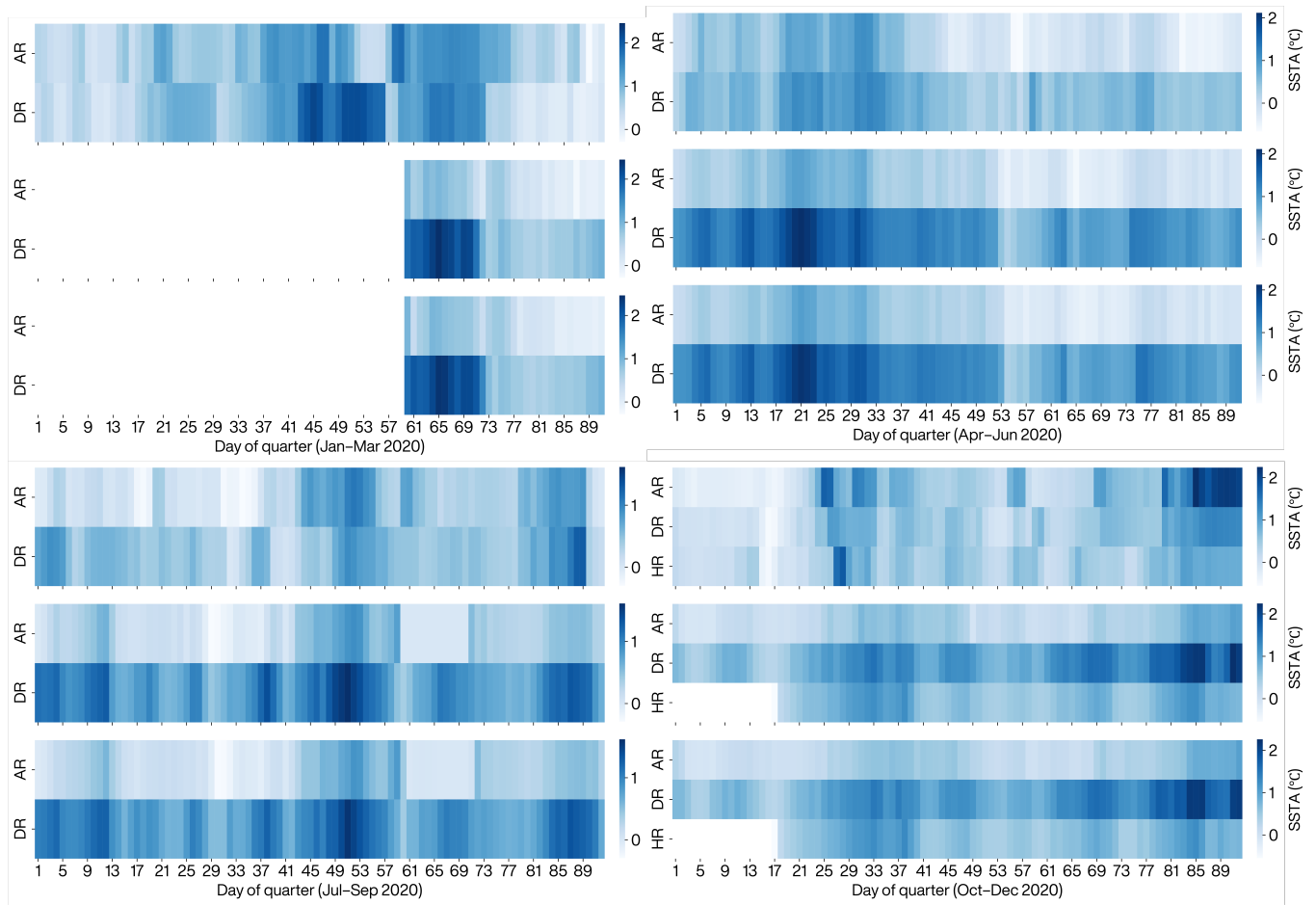
| Method          | ACC↑          | F1↑           | SEDI↑         | RMSE↓         | MSE↓          | MAE↓          |
|-----------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Hourly → Daily  | 0.8152        | 0.5719        | 0.8491        | 0.5449        | 0.2969        | 0.3270        |
| Daily → Daily   | <b>0.9231</b> | <b>0.8119</b> | <b>0.9203</b> | <b>0.2631</b> | <b>0.0692</b> | <b>0.1638</b> |
| Δ               | -0.1079       | -0.2400       | -0.0712       | +0.2818       | +0.2277       | +0.1632       |
| Daily → Weekly  | 0.8209        | 0.6087        | 0.6725        | 0.5830        | 0.3398        | 0.3768        |
| Weekly → Weekly | <b>0.8269</b> | <b>0.7187</b> | <b>0.7187</b> | <b>0.3968</b> | <b>0.1575</b> | <b>0.2701</b> |
| Δ               | -0.0060       | -0.1100       | -0.0462       | +0.1862       | +0.1823       | +0.1067       |



**Supplementary Figure 3: Station specific forecast skill of MARINA at HR.** a, Forecast skill across three temporal resolutions (hourly, daily, weekly) evaluated using SEDI, accuracy, F1 score, and RMSE (top to bottom). b, Predicted versus observed marine heatwave duration for hourly, daily, and weekly events. c, Predicted versus observed mean marine heatwave intensity for the three resolutions. d, Predicted versus observed maximum marine heatwave intensity for the three resolutions. In panels b to d, each point represents one marine heatwave event, the grey dashed line shows the one to one reference line, and the coloured solid line shows the regression fit, with the reported correlation and regression coefficients summarizing the agreement between predictions and observations.

**Supplementary Table 3: Summary of data distribution across temporal resolutions and observation sites.** The table details the total number of samples utilized for the training, validation, and test phases at hourly, daily, and weekly scales for the AR, DR, and HR weather stations.

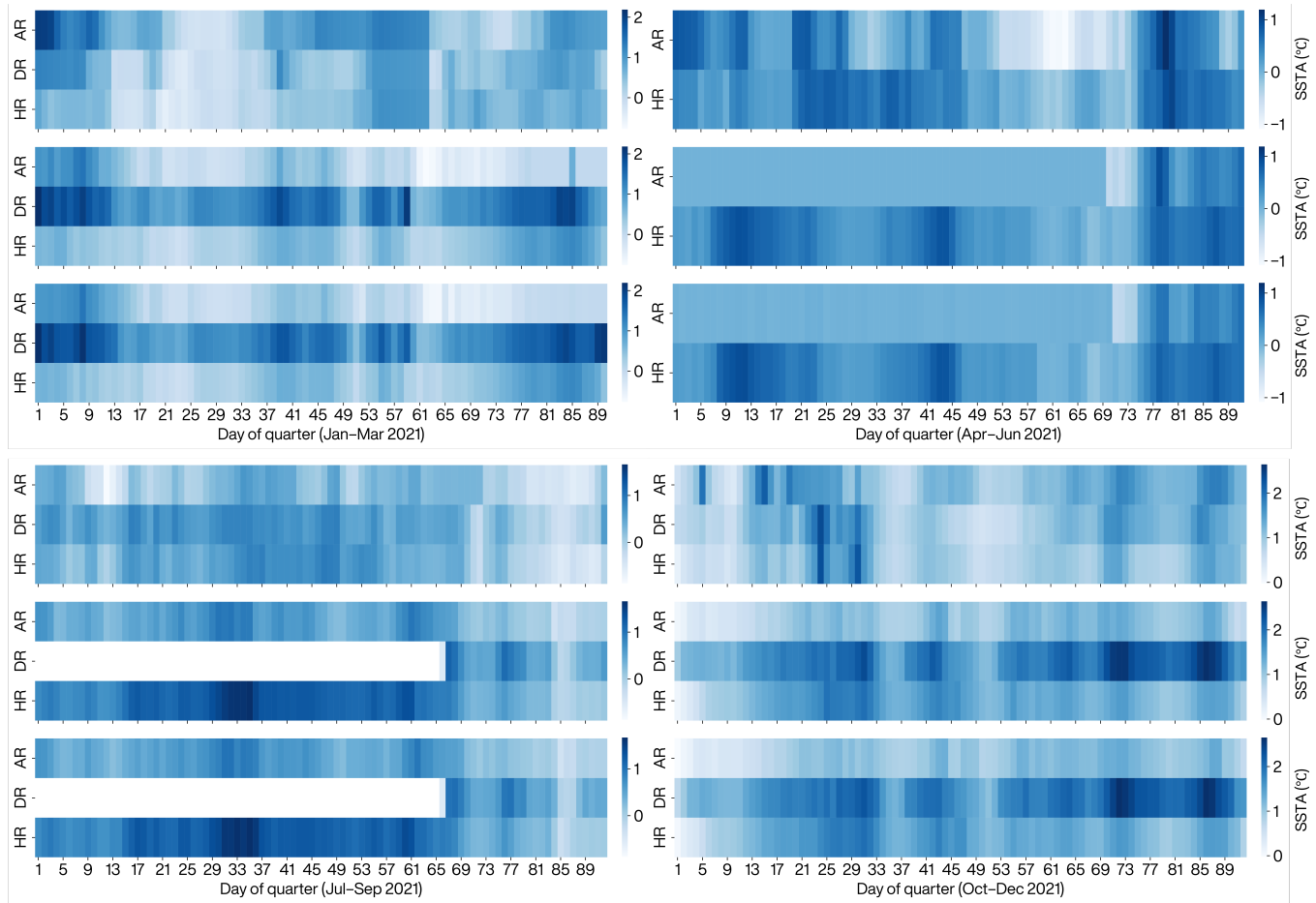
| Timescale | Weather Station | Training | Validation | Test  | ACC   | F1    |
|-----------|-----------------|----------|------------|-------|-------|-------|
| Hourly    | AR              | 161231   | 16657      | 16538 | 0.880 | 0.785 |
|           | DR              | 103233   | 7878       | 16493 | 0.932 | 0.883 |
|           | HR              | 148082   | 2626       | 7756  | 0.858 | 0.781 |
|           | AVG             | 137515   | 9054       | 13596 | 0.895 | 0.832 |
| Daily     | AR              | 9782     | 664        | 1396  | 0.923 | 0.812 |
|           | DR              | 8844     | 664        | 1233  | 0.929 | 0.875 |
|           | HR              | 8110     | 386        | 1165  | 0.944 | 0.913 |
|           | AVG             | 8912     | 571        | 1265  | 0.932 | 0.866 |
| Weekly    | AR              | 833      | 126        | 249   | 0.827 | 0.719 |
|           | DR              | 833      | 126        | 249   | 0.827 | 0.654 |
|           | HR              | 833      | 126        | 249   | 0.851 | 0.716 |
|           | AVG             | 833      | 126        | 249   | 0.835 | 0.696 |



**Supplementary Figure 4: Seasonal anomaly patterns across four quarters of 2020 at three reef stations (AR, DR, and HR).** The x-axis represents the day of quarter (spanning sequential calendar days for each period), while the y-axis denotes the reef locations. Each subplot presents a comparison between global observations (top), local observations (middle), and local predictions (bottom). The color bar indicates the sea surface temperature anomaly (SSTA) magnitude in degrees Celsius, with darker shades representing more intense positive anomalies. White areas denote missing data.

**Supplementary Table 4: Comparative performance of MARINA under limited and full-data training regimes.** Results demonstrate the model's predictive skill across four metrics (ACC, F1, SEDI, and RMSE) when trained on a reduced data ratio (40% for AR and DR; 50% for HR) compared to the full dataset. The stability of the results across configurations indicates the model's robustness in data-scarce environments. AVG means macro average.

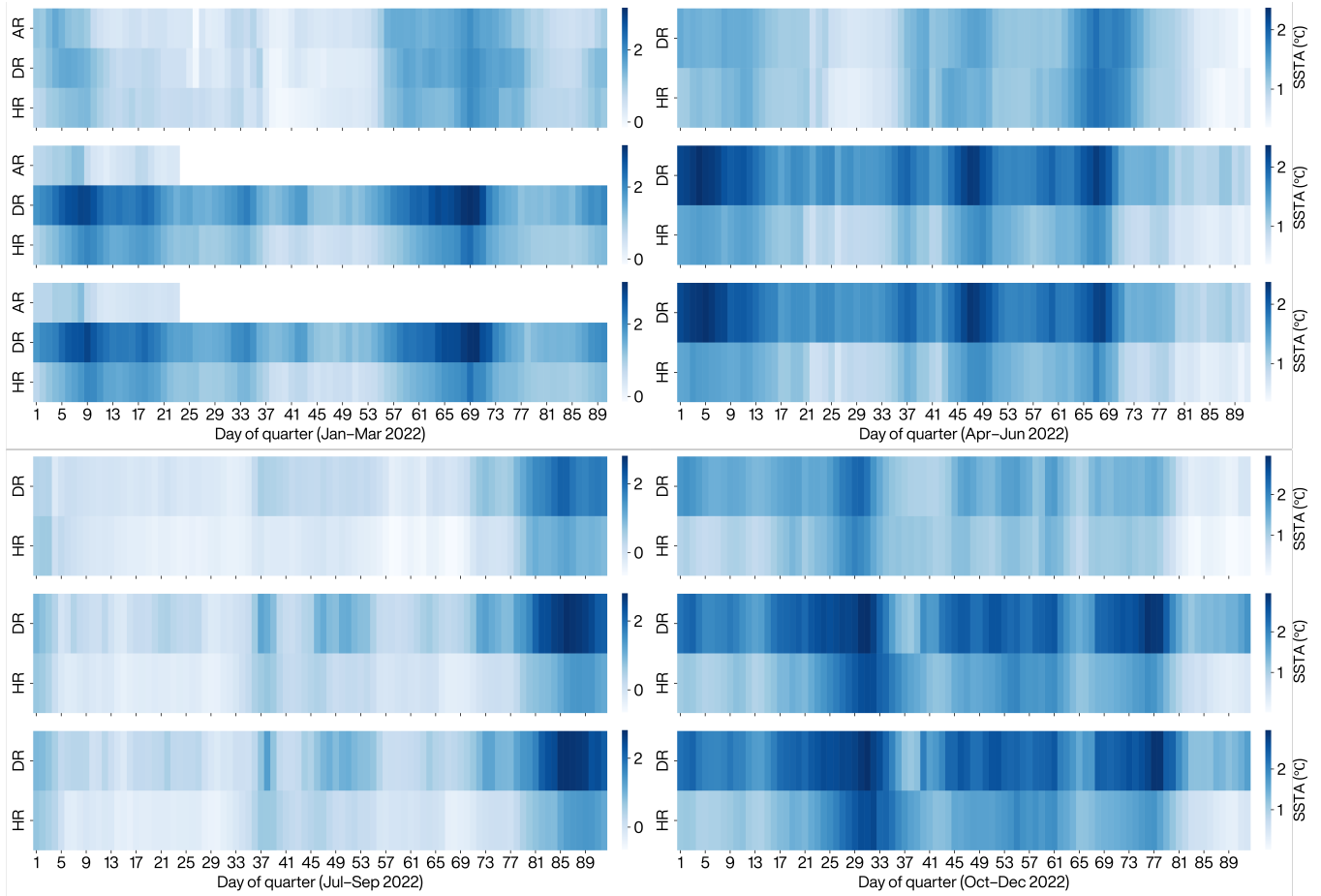
| Station | Model      | ACC   | F1    | SEDI  | RMSE  |
|---------|------------|-------|-------|-------|-------|
| AR      | 40% data   | 0.866 | 0.767 | 0.908 | 0.313 |
|         | Full data  | 0.880 | 0.785 | 0.915 | 0.309 |
| DR      | 40% data   | 0.906 | 0.845 | 0.942 | 0.234 |
|         | Full data  | 0.932 | 0.883 | 0.957 | 0.246 |
| HR      | 50% data   | 0.859 | 0.783 | 0.930 | 0.169 |
|         | Full data  | 0.858 | 0.781 | 0.929 | 0.175 |
| AVG     | Few-sample | 0.877 | 0.798 | 0.927 | 0.239 |
|         | Full data  | 0.895 | 0.832 | 0.943 | 0.211 |



**Supplementary Figure 5: Seasonal anomaly patterns across four quarters of 2021 at three reef stations (AR, DR, and HR).** The x-axis represents the day of quarter (spanning sequential calendar days for each period), while the y-axis denotes the reef locations. Each subplot presents a comparison between global observations (top), local observations (middle), and local predictions (bottom). The color bar indicates the sea surface temperature anomaly (SSTA) magnitude in degrees Celsius, with darker shades representing more intense positive anomalies. White areas denote missing data.

**Supplementary Table 5: Description of predefined experimental operators.** The first column lists the operators used in our experiments, while the second column provides their corresponding descriptions. These operators can describe various atmospheric processes. For example,  $\max(\text{air temperature}, \text{air pressure})$  can be used to determine which factor has a greater impact on current SST anomalies or MHW events. The formal definitions of these operators follow those described in Yu et al.<sup>1</sup>.

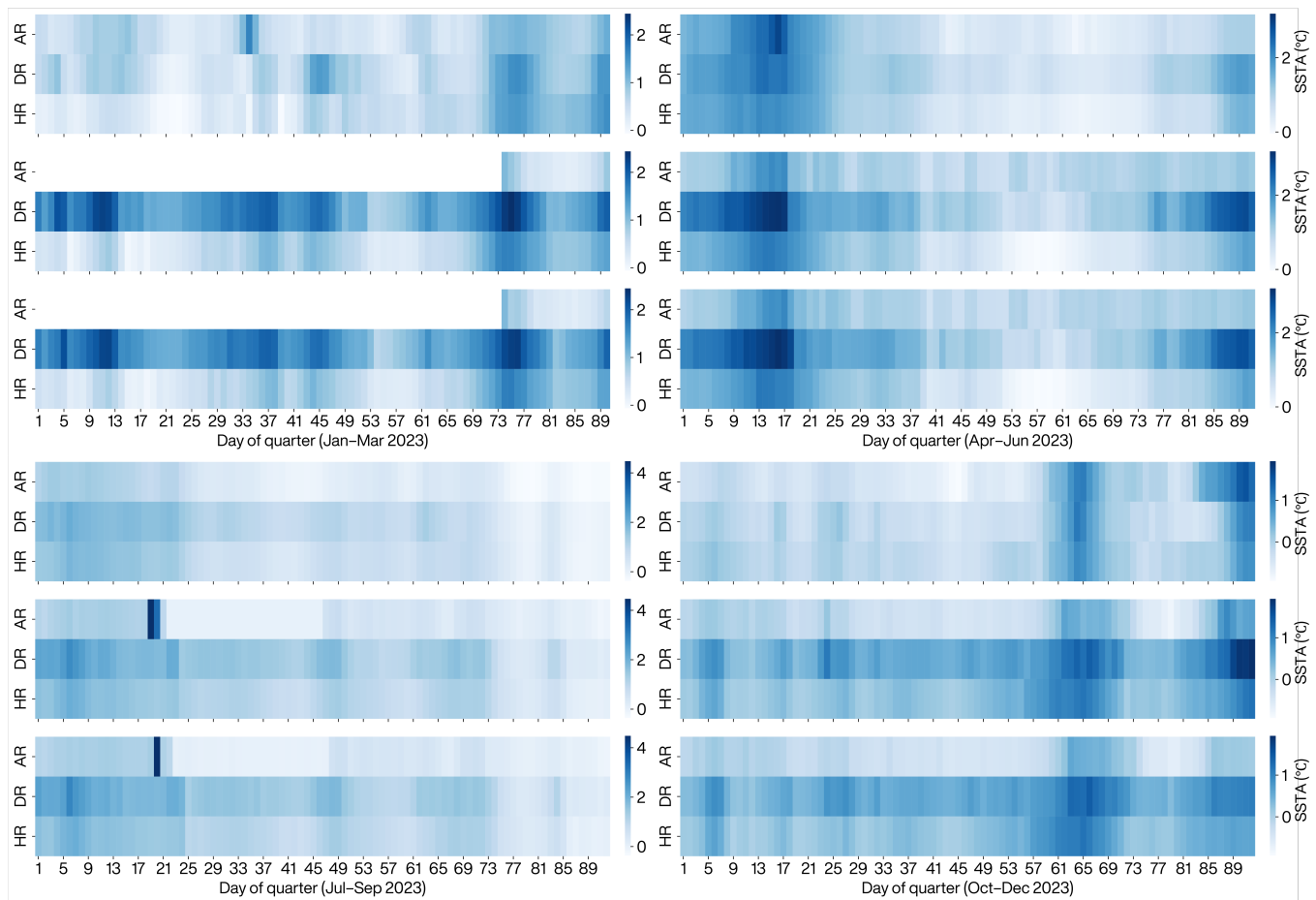
| Operator                         | Descriptions                                                                                               |
|----------------------------------|------------------------------------------------------------------------------------------------------------|
| Abs(x)                           | The absolute value $ x $ .                                                                                 |
| Log(x)                           | Natural logarithmic function $\log(x)$ .                                                                   |
| $x + y, x - y, x \cdot y, x/y$   | Arithmetic operators.                                                                                      |
| Greater(x, y), Less(x, y)        | The larger/smaller one of the two values.                                                                  |
| Ref(x, t)                        | The expression x evaluated at t days before the current day.                                               |
| Mean(x, t), Med(x, t), Sum(x, t) | The mean/median/sum value of the expression x evaluated on the recent t days.                              |
| Std(x, t), Var(x, t)             | The standard deviation/variance of the expression x evaluated on recent t days.                            |
| Max(x, t), Min(x, t)             | The maximum/minimum value of the expression x evaluated on the recent t days.                              |
| Mad(x, t)                        | The mean absolute deviation $E[ x - E[x] ]$ of the expression x evaluated on the recent t days.            |
| Delta(x, t)                      | The relative difference of x compared to t days ago, $x - \text{Ref}(x, t)$ .                              |
| WMA(x, t), EMA(x, t)             | Weighted moving average and exponential moving average of the expression x evaluated on the recent t days. |
| Cov(x, y, t)                     | The covariance between two time series x and y in the recent t days.                                       |
| Corr(x, y, t)                    | The Pearson's correlation coefficient between two time series x and y in recent t days.                    |



**Supplementary Figure 6: Seasonal anomaly patterns across four quarters of 2022 at three reef stations (AR, DR, and HR).** The x-axis represents the day of quarter (spanning sequential calendar days for each period), while the y-axis denotes the reef locations. Each subplot presents a comparison between global observations (top), local observations (middle), and local predictions (bottom). The color bar indicates the sea surface temperature anomaly (SSTA) magnitude in degrees Celsius, with darker shades representing more intense positive anomalies. White areas denote missing data.

**Supplementary Table 6:** Quantitative performance comparison between MARINA and the Diurnal Cycle baseline across three representative reef stations. MARINA demonstrates significantly lower error rates across all stations and metrics,

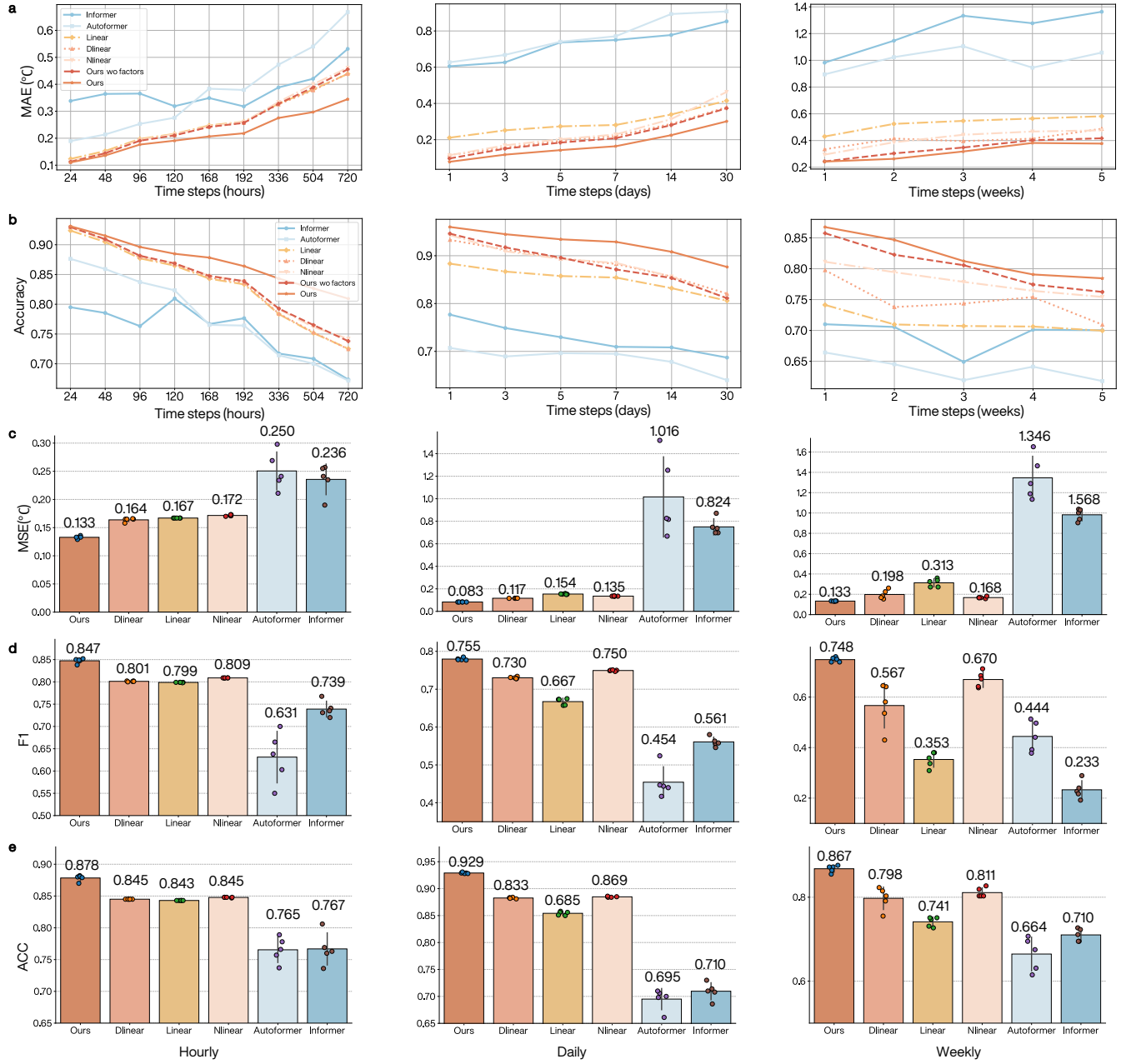
| Station | Model        | RMSE ↓        | MSE ↓         | MAE ↓         |
|---------|--------------|---------------|---------------|---------------|
| AR      | MARINA       | <b>0.3087</b> | <b>0.0953</b> | <b>0.1784</b> |
|         | Durnal Cycle | 1.7225        | 2.9670        | 1.5167        |
| DR      | MARINA       | <b>0.2463</b> | <b>0.0607</b> | <b>0.1691</b> |
|         | Durnal Cycle | 2.1383        | 4.5725        | 1.8800        |
| HR      | MARINA       | <b>0.1753</b> | <b>0.0307</b> | <b>0.1019</b> |
|         | Durnal Cycle | 2.1736        | 4.7244        | 1.8979        |
| AVG     | MARINA       | <b>0.2434</b> | <b>0.0622</b> | <b>0.1498</b> |
|         | Durnal Cycle | 2.0115        | 4.0880        | 1.7649        |



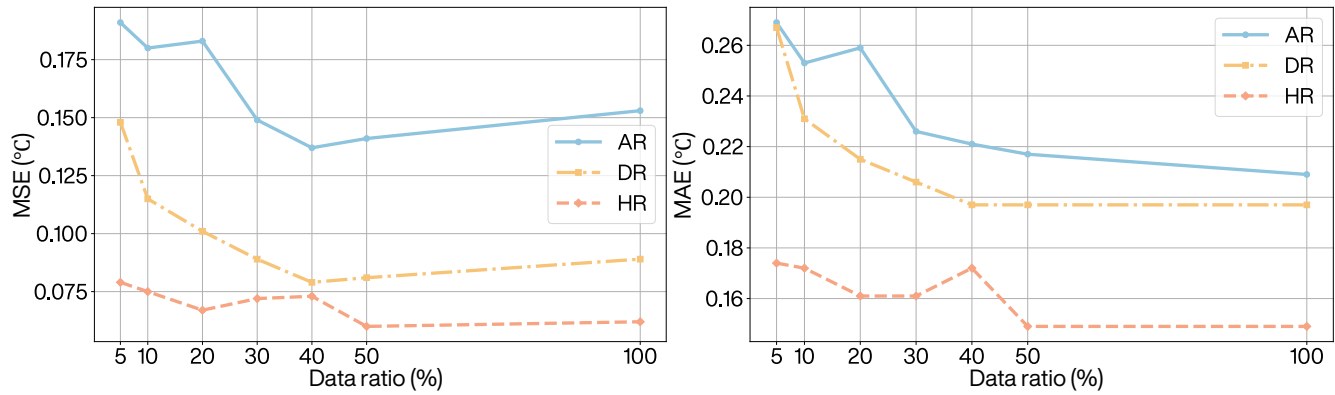
**Supplementary Figure 7: Seasonal anomaly patterns across four quarters of 2023 at three reef stations (AR, DR, and HR).** The x-axis represents the day of quarter (spanning sequential calendar days for each period), while the y-axis denotes the reef locations. Each subplot presents a comparison between global observations (top), local observations (middle), and local predictions (bottom). The color bar indicates the sea surface temperature anomaly (SSTA) magnitude in degrees Celsius, with darker shades representing more intense positive anomalies. White areas denote missing data.

**Supplementary Table 7:** Publicly accessible high-frequency (hourly/daily) global data sources compatible with the MARINA framework for out-of-region deployment.

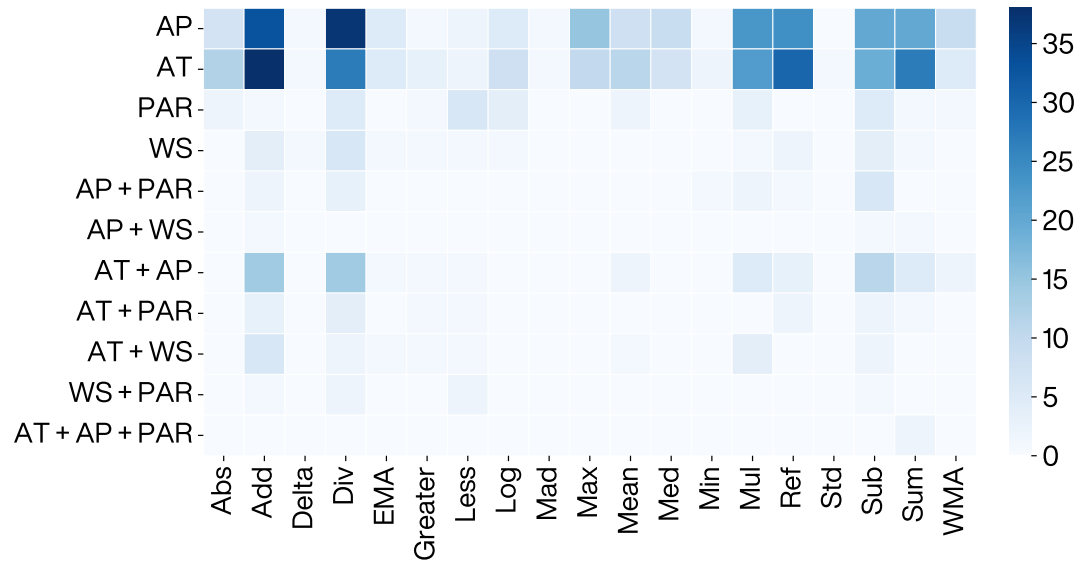
| Source          | Type            | Access Link                                                                                                                                                                                                                                                                     |
|-----------------|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| NOAA NDBC       | In-situ Buoys   | <a href="https://www.ndbc.noaa.gov">https://www.ndbc.noaa.gov</a>                                                                                                                                                                                                               |
| NOAA CO-OPS     | Coastal Sensors | <a href="https://catalog.data.gov/dataset/coastal-meteorological-and-water-temperature-data-from-the-noaa-center-for-operational-oceanogr">https://catalog.data.gov/dataset/coastal-meteorological-and-water-temperature-data-from-the-noaa-center-for-operational-oceanogr</a> |
| ERA5 Reanalysis | Global Grid     | <a href="https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-timeseries">https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-timeseries</a>                                                                                           |
| NOAA OISST      | Daily Global    | <a href="https://downloads.psl.noaa.gov/Datasets/noaa.oisst.v2.highres">https://downloads.psl.noaa.gov/Datasets/noaa.oisst.v2.highres</a>                                                                                                                                       |



**Supplementary Figure 8: Additional comparison of forecasting performance across models.** Each column presents results for a different temporal resolution: hourly (left), daily (middle), and weekly (right) SST prediction. **a–b**, MAE and accuracy evaluated over multiple forecast horizons. **c–e**, Bar plots comparing MSE, F1 score, and ACC across models, with scatter points indicating results from five independent runs. Across all settings and metrics, our model consistently achieves the best or competitive performance, demonstrating robustness across timescales and prediction lengths.



**Supplementary Figure 9: Sensitivity of MARINA model performance to training data volume.** MSE and MAE are plotted as a function of the training data ratio (ranging from 5% to 100%) for the AR, DR, and HR stations. The observed performance plateau at approximately 40–50% data volume suggests that the model achieves high predictive skill without requiring exhaustive high-frequency datasets.



**Supplementary Figure 10: The full occurrence heatmap between meteorological factors and operators.** Darker colors indicate a higher influence ratio of these factor pairs on SST. Air temperature and air pressure emerge as the most important factors. AP: Air pressure, AT: Air temperature, WS: Wind speed, PAR: photosynthetically active radiation.



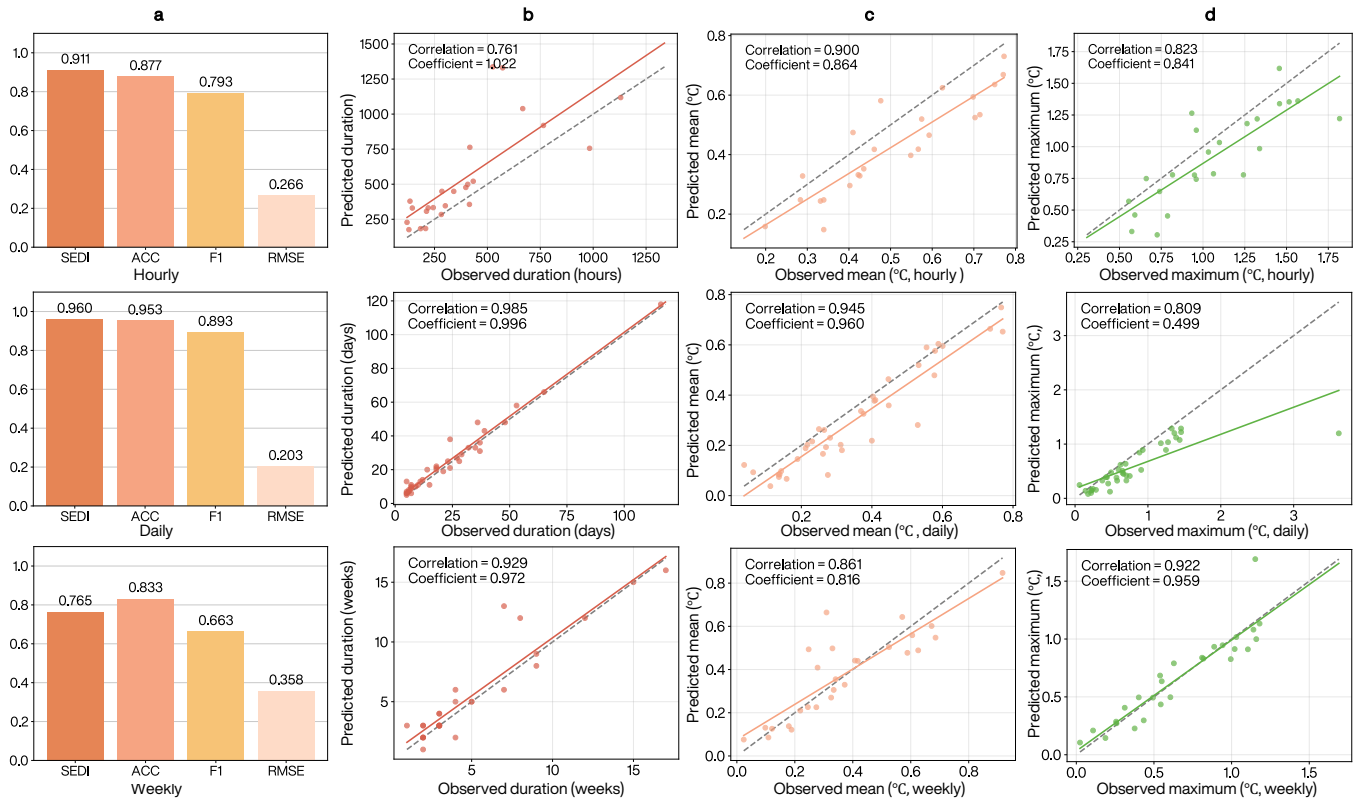
## GPT4 Prompt Template for Marina: User Prompt Details

```

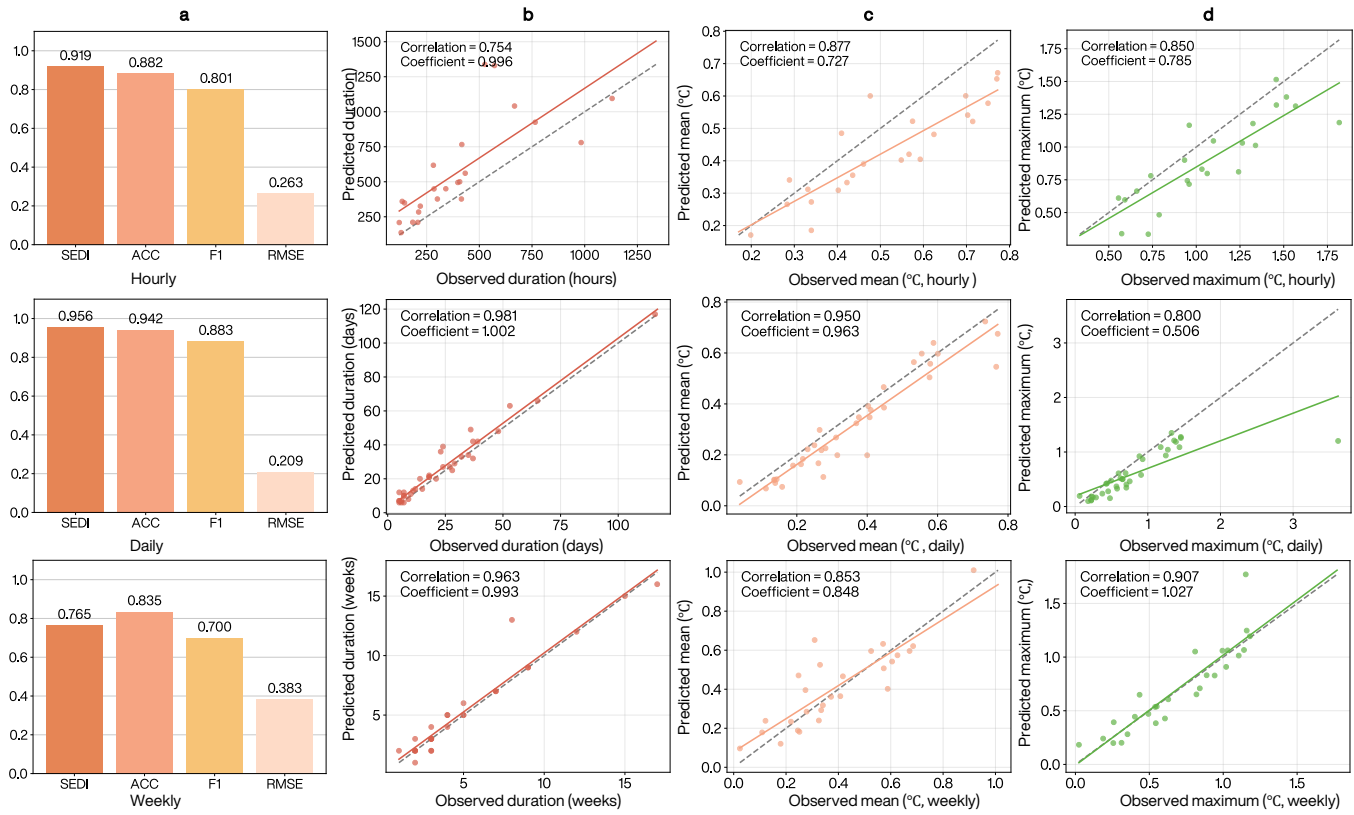
# system_prompt
{"role": "system",
"content": "You are an expert in the field of neural architecture search."},
# user_prompt
{"role": "user",
# task description
"content": "Your task is to assist me in selecting the best parameters for a given model architecture, which includes some undefined layers and available operations. The model will be trained and tested on MT-MHW, and your objective will be to maximize the model's performance on MT-MHW. MT-MHW is the SST (Sea Surface Temperature) and MHW (Marine heatwave) dataset. MT-MHW includes key meteorological variables that govern MHW evolution, providing a comprehensive basis for research on multiple timescales MHW forecasting. Different timescale has different characteristics. Hourly data are noisy with fine-grained patterns; daily data show clearer seasonality and moderate trends; weekly data are smoother, highlighting long-term dynamics. I will tell you the frequency of the current dataset. Please select the proper parameters according to the frequency.",
# Model Architecture
"model_definition": {
"class_name": "Model",
"components": [
"series_decomp(kernel_size_1)",
"dropout_layer(dropout)",
"predict_linear_1(seq_len -> seq_len + pred_len)",
"block1(enc_in -> d_model1 -> d_model2 -> d_model3, kernel_size_1, layer_num_block1)",
.....
],
},
# Search Space
"undefined_parameters": {
"layer_num_1": ["0", "1", "2", "3", "4", "5"],
"layer_num_2": ["0", "1", "2", "3", "4", "5"],
"d_model1": ["64", "128", "256", "512", "768", "1024"],
"dropout": ["0.0", "0.1", "0.2", "0.3", "0.4", "0.5"],
"kernel_size_1": ["3", "5", "7", "9", "21", "23", "25"],
.....},
"instruction": "maximize the model's performance on MT-MHW, please provide me with your suggested operation for the undefined parameters only. Your response should be an operation ID list for the undefined parameters. For example: [1, 2, ..., 0] means we use operation 1 for layer_num_1, operation 2 for layer_num_2, ..., operation 0 for kernel_size_1.
# performance history (Recent Iterations)
Here are some experimental results that you can use as a reference: [3, 2, 4, 3, 2, .....] gives a MSE of 0.10, a MAE of 0.18, [4, 3, 5, 4, 3, .....] gives a MSE of 0.11, a MAE of 0.18, [2, 4, 3, 2, 1, .....] gives a MSE of 0.10. Please suggest another parameter list that can improve the model's performance on SST and MHW dataset beyond the experimental results provided above. You can provide more architecture options. current dataset is {hourly, daily, or weekly} datasets, please select parameters accounting for time series characteristic. Please do not include anything other than the operation ID list in your response."}

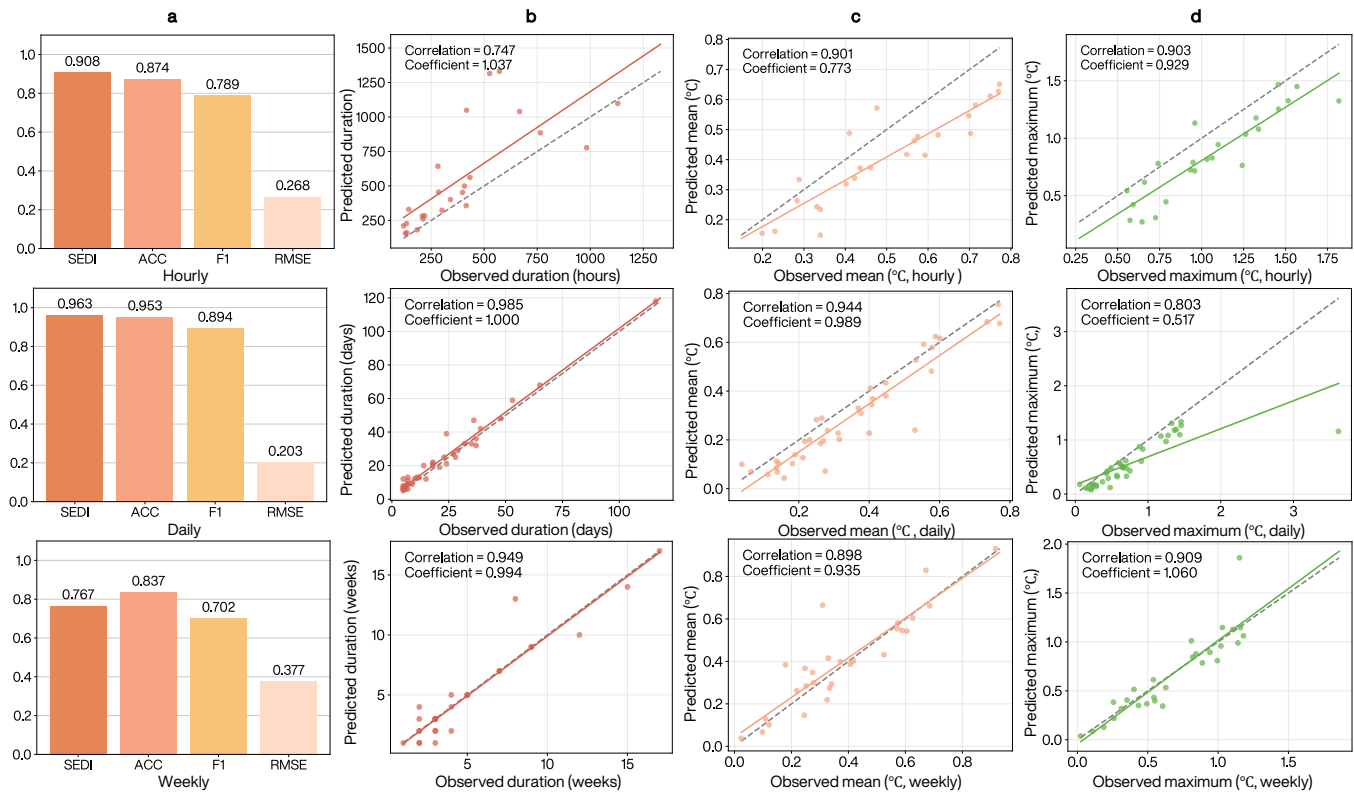
```

**Supplementary Figure 11: The detailed prompt for LLM-enhanced neural network design.** During training, MARINA leverages GPT-4 to automatically design the architecture of the dual-branch model. For each temporal resolution, MARINA provides GPT-4 with the frequency of the current dataset, enabling it to select appropriate architectural parameters based on the dataset's characteristics.

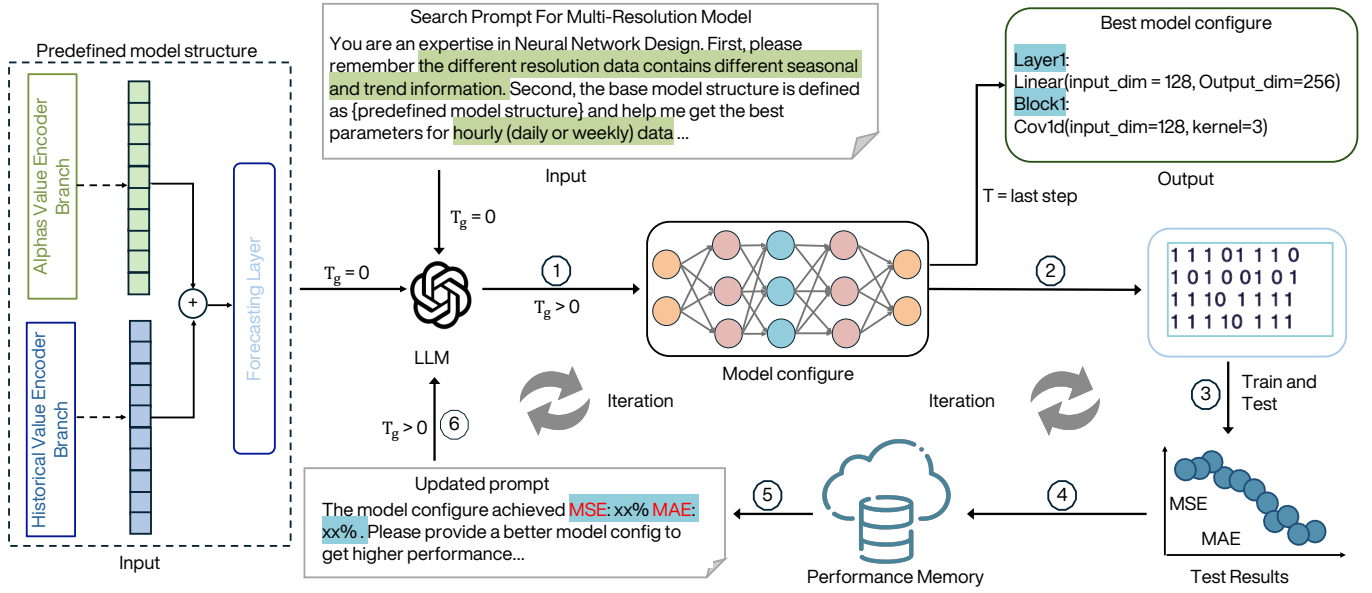


**Supplementary Figure 12: Multi-resolution forecast skill of MARINA with U/V current for MHW events.** **a**, Summary of event classification and regression performance based on the micro-mean across three reef stations, evaluated in SEDI, ACC, F1 and RMSE. MARINA performs strongly across all resolutions, with the strong SEDI value  $> 0.9$  in hourly and daily resolution. **b–d**, Predicted versus observed marine heatwave duration (b), mean intensity (c), and maximum intensity (d) at each temporal resolution. Each dot represents a single *true-positive* (successfully predicted) observed MHW event. Pearson correlation coefficients (correlation  $r$ ) quantify agreement in temporal variability, while regression slopes (coefficient  $\alpha$ ) indicate scale fidelity. Dotted lines denote the 1:1 reference ( $y = x$ ). Solid lines represent the fitted linear regression. High correlations and coefficients are maintained across all resolutions (average  $r > 0.85$ ,  $\alpha > 0.85$ ), confirming the model's capacity to capture critical temporal attributes of MHW events. Missed events are not shown in panels b, c, and d but are accounted for in the event based metrics reported in panel a.





**Supplementary Figure 14: Multi-resolution forecast skill of MARINA with OHC for MHW events.** **a**, Summary of event classification and regression performance based on the micro-mean across three reef stations, evaluated in SEDI, ACC, F1 and RMSE. MARINA performs strongly across all resolutions, with the strong SEDI value  $> 0.9$  in hourly and daily resolution. **b–d**, Predicted versus observed marine heatwave duration (b), mean intensity (c), and maximum intensity (d) at each temporal resolution. Each dot represents a single *true-positive* (successfully predicted) observed MHW event. Pearson correlation coefficients (correlation  $r$ ) quantify agreement in temporal variability, while regression slopes (coefficient  $\alpha$ ) indicate scale fidelity. Dotted lines denote the 1:1 reference ( $y = x$ ). Solid lines represent the fitted linear regression. High correlations and coefficients are maintained across all resolutions (average  $r > 0.85$ ,  $\alpha > 0.85$ ), confirming the model's capacity to capture critical temporal attributes of MHW events. Missed events are not shown in panels b, c, and d but are accounted for in the event based metrics reported in panel a.



**Supplementary Figure 15: The detail pipeline of LLM enhanced neural network design.** At  $T_g = 0$ , GPT-4 is prompted with a predefined search prompt and initial model structure. GPT proposes a model configuration and we implement and run the corresponding model code to evaluate its performance (e.g., MSE, MAE, SEDI). Based on the evaluation results, we update the prompt to guide GPT in refining the model configuration. This process repeats until the model's performance stabilizes. This schematic illustration is conceptually informed by the LLM-enhanced neural network search framework established by Zheng et al.<sup>2</sup>

**Input:** Raw observations from candidate stations  $\mathcal{S}$  over the baseline period (1980–2023). Station missing threshold  $\beta$  and variable missing threshold  $\alpha$ .

**Output:** Multi-resolution (hourly/daily/weekly) dataset with resolution-specific MHW labels.

```

1: Station Selection.
2: for each station  $s \in \mathcal{S}$  do
3:   Compute missing rate  $r_s$  over the baseline period.
4:   if  $r_s > \beta$  then
5:     Remove  $s$  from  $\mathcal{S}$ .
6:   end if
7: end for
8: Variable Filtering.
9: for each variable  $x \in \mathcal{V}$  do
10:   Compute missing rate  $m_x$  across retained stations over the baseline period.
11:   if  $m_x > \alpha$  then
12:     Remove  $x$  from  $\mathcal{V}$ .
13:   end if
14: end for
15: Data Standardization Across Resolutions.
    Hourly resampling
16: for each station  $s \in \mathcal{S}$  do
17:   for each variable  $x \in \mathcal{V}$  do
18:     Resample high-resolution observations to hourly intervals via mean aggregation.
19:   end for
20: end for
    Cross-resolution consistency and gap exclusion
21: for each resolution  $r \in \{\text{hourly, daily, weekly}\}$  do
22:   for each station  $s \in \mathcal{S}$  do
23:     Apply identical quality control and filtering procedures.
24:     if any continuous missing span  $> 30$  days then
25:       Discard the corresponding time span.
26:     end if
27:     Impute remaining missing values.
28:   end for
29: end for
30: MHW Labeling (Resolution-specific).
31: for each resolution  $r \in \{\text{hourly, daily, weekly}\}$  do
32:   Estimate the resolution-specific climatology  $M_r$  and the 90th-percentile threshold  $\tau_r$ .
33:   Select resolution-specific parameters (e.g., minimum duration and maximum gap length).
34:   Detect MHW events and generate labels using Hobday et al. with  $M_r$  and  $\tau_r$ .
35: end for

```

**Supplementary Algorithm 1:** MT-MHW Data Processing and Labeling Pipeline

## REFERENCES

- [1] Shuo Yu, Hongyan Xue, Xiang Ao, Feiyang Pan, Jia He, Dandan Tu, and Qing He. Generating synergistic formulaic alpha collections via reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, page 11, New York, NY, USA, 2023. ACM.
- [2] Ming Zheng, Xiaoyu Su, Shaofu You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. Can gpt-4 perform neural architecture search? *arXiv preprint arXiv:2304.10970*, 2023.