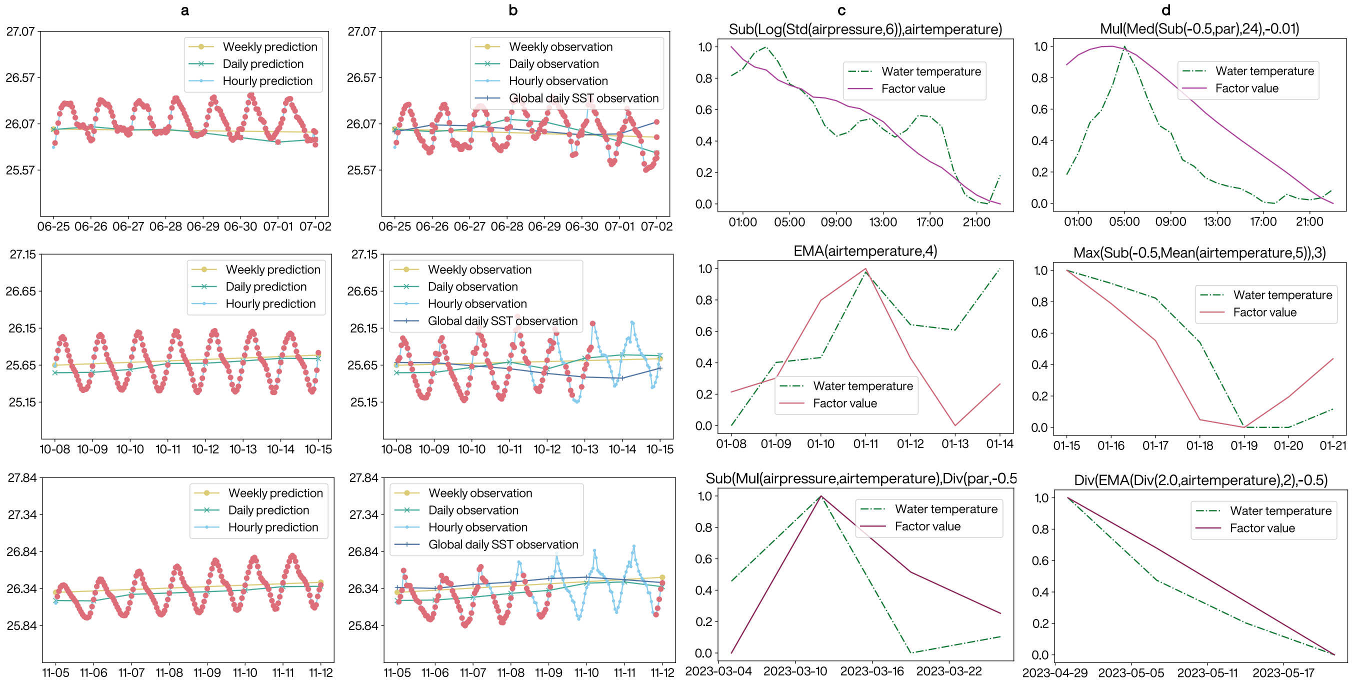
Supplementary Information for
A Lightweight Physics-Aware Framework for Multi-Scale Marine Heatwaves Forecasting

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

[1]*Central South University, Changsha, China*
[2]*University of Sydney, Sydney, Australia*
[3]*Fudan University, Shanghai, China*
[4]*The Chinese University of Hong Kong, Hong Kong, China*
[5]*Shanghai Jiao Tong University, Shanghai, China*
[6]*Tongji University, Shanghai, China*
[7]*SenseTime Research, Beijing, China*

The PDF file includes:

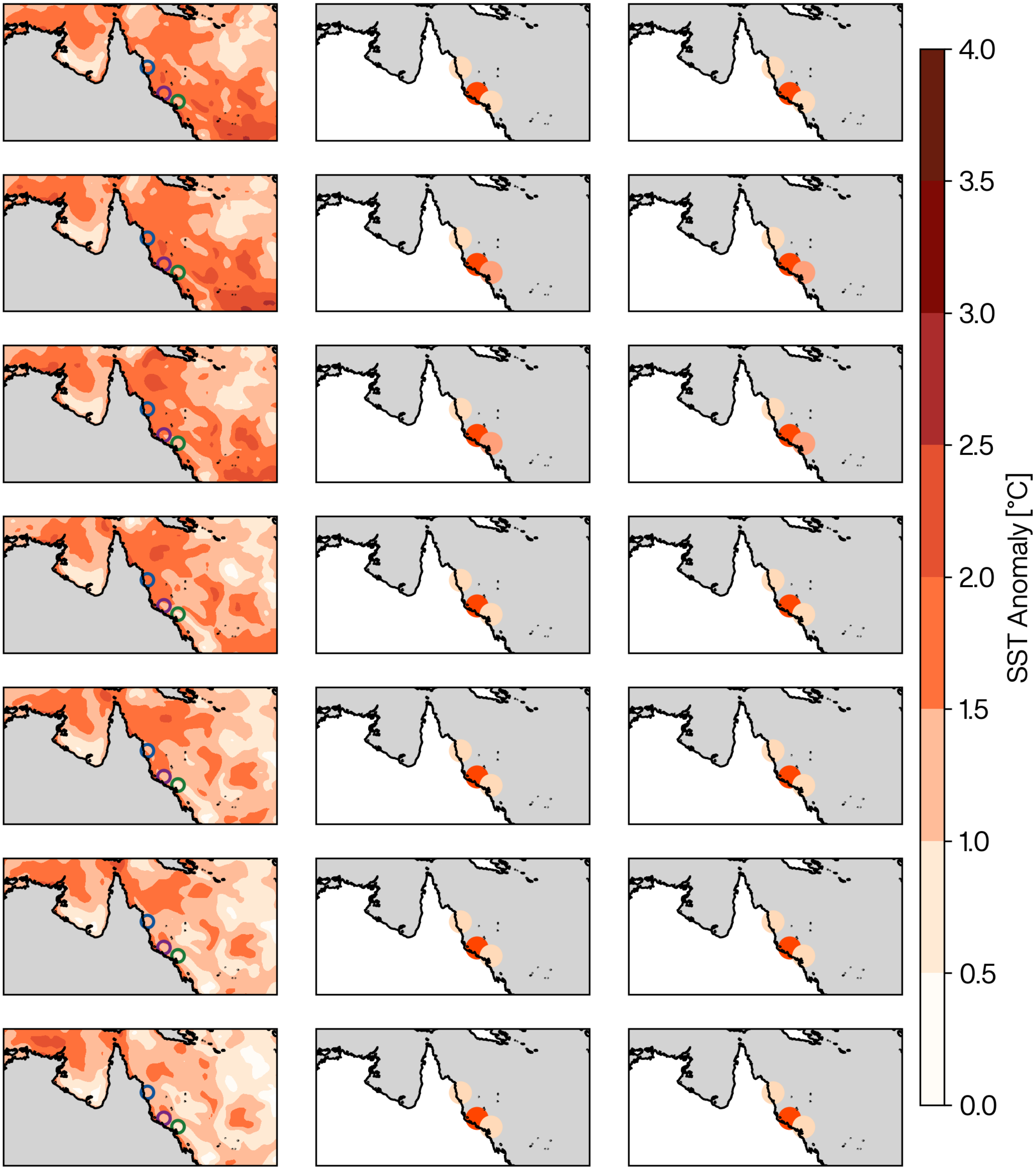
[*] These authors contributed equally.



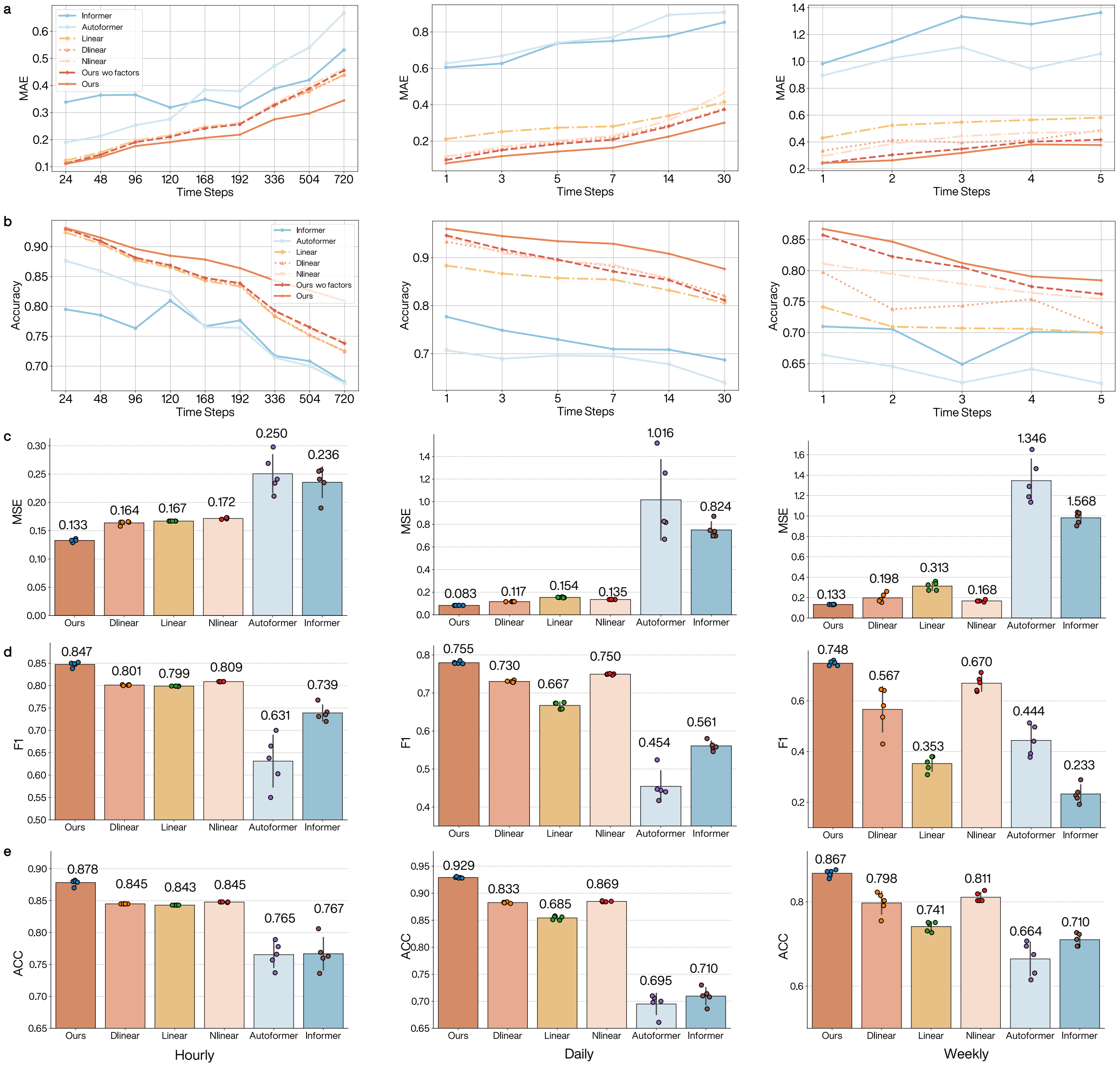
Supplementary Figure 1. **Additional performance visualizations of MARINA across temporal resolutions. a-b**, One-week forecasts at hourly, daily and weekly resolutions are compared with global daily observations from 2020 to 2023. The comparison demonstrates the accuracy of our model at different time resolutions, with finer-grained hourly and daily predictions closely aligning with the global observations, while weekly predictions capture broader trends. **c-d**, the comparison between three types of the value of interaction factor and the corresponding SST value. The title of each sub-figure represents the generated interaction factor formula used for prediction. The overall predicted trends show strong consistency with the ground truth across hourly, daily and weekly resolutions,demonstrating the effectiveness of our interaction factor-based forecasting method in capturing SST variations over different time scales.

Supplementary Table 1. **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(air temperature, air pressure)* can be used to determine which factor has a greater impact on current SST anomalies or MHW events.

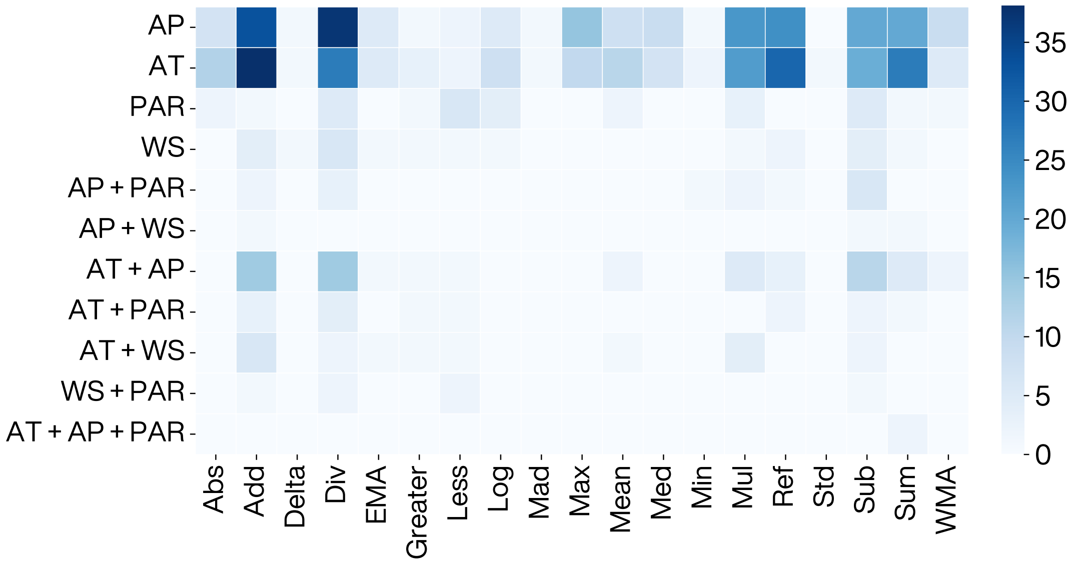
Operator	**Descriptions**
Abs(x)	The absolute value $|x|$.
Log(x)	Natural logarithmic function $log(x)$.
$x + y, x - y, x \cdot y, x/y$	Arithmetic operators.
Greater(x, y), Less(x, y)	The larger/smaller one of the two values.
Ref(x, t)	The expression x evaluated at t days before the current day.
Mean(x, 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 – 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.



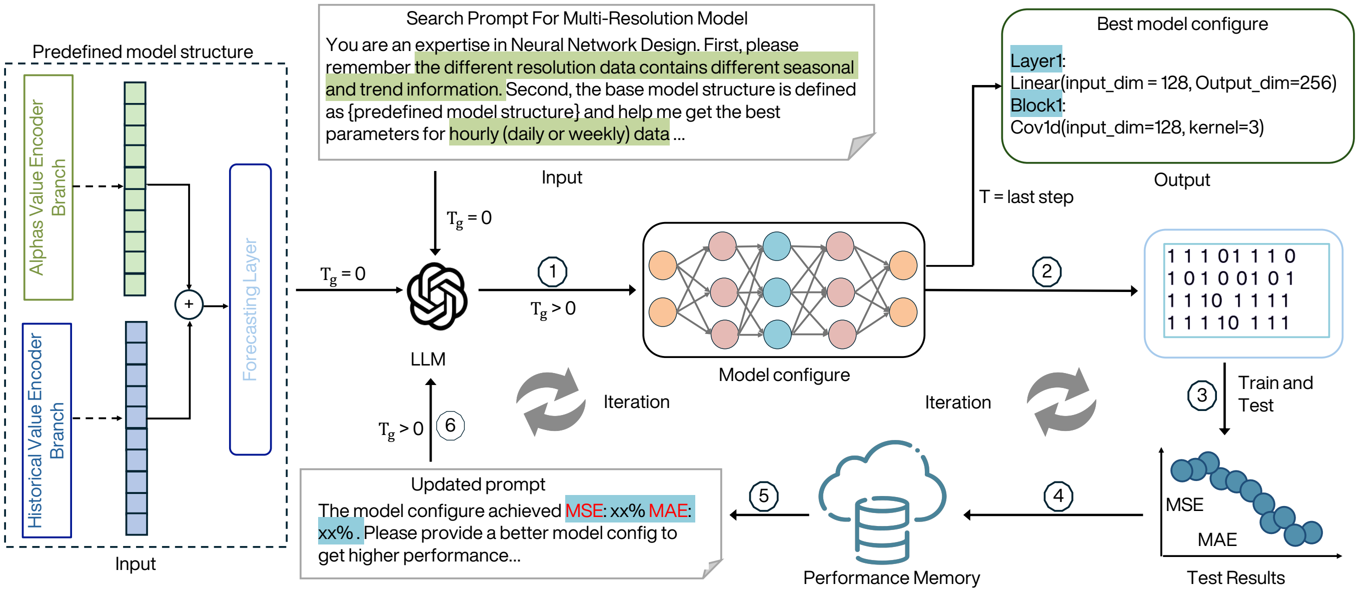
Supplementary Figure 2. **Comparison between global observations and the observations and predictions from three local weather stations.** The plot compares model predictions and local observations with global SST data over the period from 10 to 16 December 2021. The first column shows the global observations, where the blue, orange, and green points denote Agincourt Reef, Davies Reef, and Hardy Reef, respectively. The second and third columns display the local observations and corresponding model predictions from the three weather stations. Our model achieves close alignment with both global and locally observed SST patterns.



Supplementary Figure 3. **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 4. **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.


Predefined model structure
Alphas Value Encoder Branch
Historical Value Encoder Branch
+
Forecasting Layer
Input
Search Prompt For Multi-Resolution Model

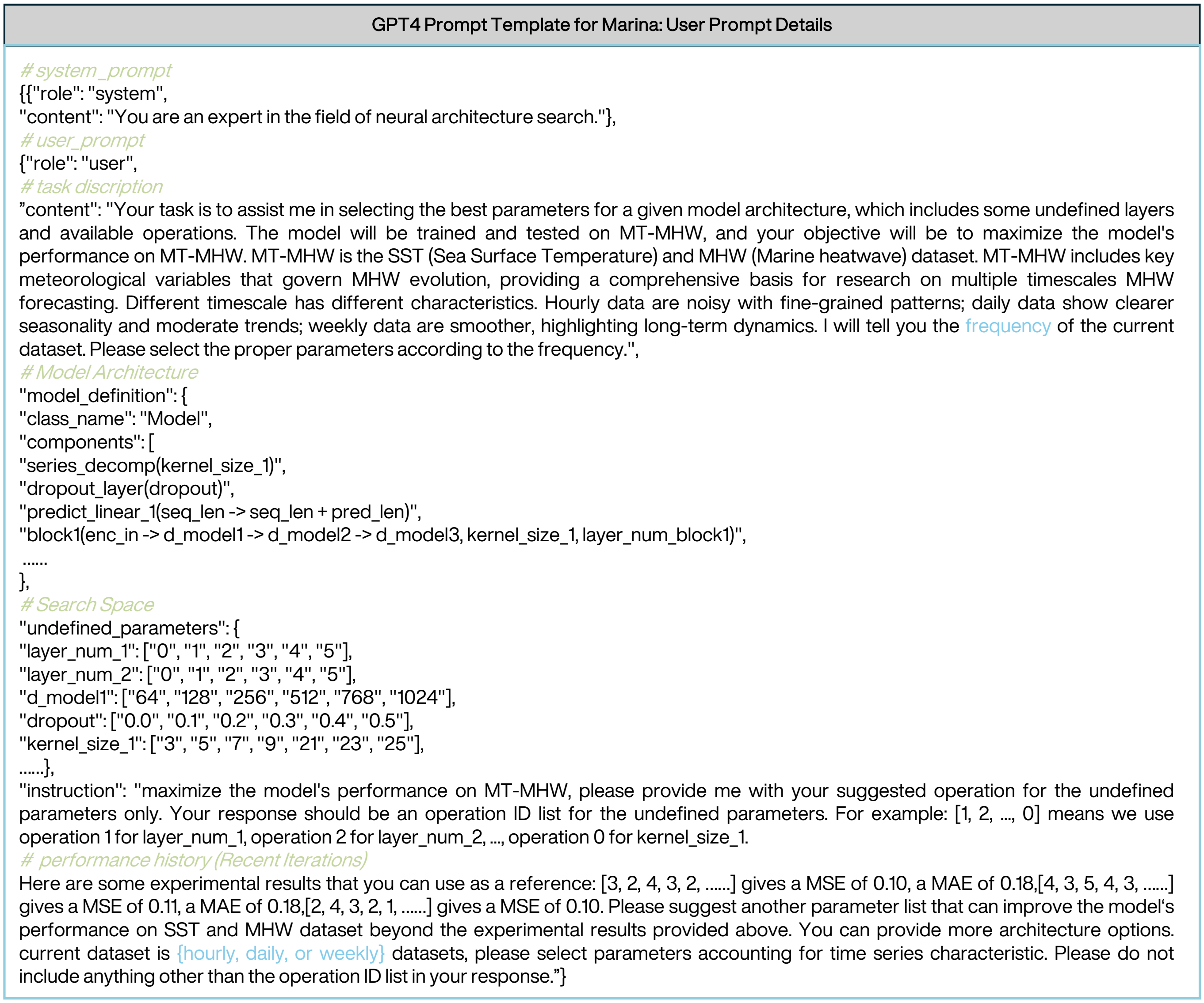
You are an expertise in Neural Network Design. First, please remember the different resolution data contains different seasonal and trend information. Second, the base model structure is defined as {predefined model structure} and help me get the best parameters for hourly (daily or weekly) data ...

Input
$T_g = 0$
$T_g = 0$
LLM
1
$T_g > 0$
Model configure
T = last step
Best model configure

Layer1:
Linear(input_dim = 128, Output_dim=256)
Block1:
Cov1d(input_dim=128, kernel=3)

The model configure achieved MSE: xx% MAE: xx% . Please provide a better model config to get higher performance...

Supplementary Figure 5. **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.



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 discription
”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 6. **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.