A Multiple-Location Modeling Scheme for Physics-Regularized Networks: Recurrent Forecasting of Fixed-Location Buoy Observations

Elias Sandner
CoDiS-Lab ISDS
Graz University of Technology
Graz, Austria
email: sandner@student.tugraz.at

Austin Schmidt
GulfSCEI
University of New Orleans
New Orleans, United States
email: sbaustin@uno.edu

Pujan Pokhrel
GulfSCEI
University of New Orleans
New Orleans, United States
email: ppokhre1@uno.edu

Elias Ioup
Center for Geospatial Sciences
Naval Research Laboratory
Mississippi, United States
email: elias.ioup@nrlssc.navy.mil

David Dobson
Center for Geospatial Sciences
Naval Research Laboratory
Mississippi, United States
email: david.dobson@nrlssc.navy.mil

Christian Guetl
CoDiS-Lab ISDS
Graz University of Technology
Graz, Austria
email: c.guetl@tugraz.at

Mahdi Abdelguerfi
GulfSCEI
University of New Orleans
New Orleans, United States
email: gulfsceidirector@uno.edu

***Abstract*—Reliable oceanic and climate analysis depend on high-quality sensor readings, yet these systems commonly encounter significant sensor limitations, leading to missing data. Addressing this issue is critical for ensuring accurate forecasts and analyses. In this work, the data gap problem is studied by developing physics-regularized machine learning models with multiple-location modeling to forecast missing sensor data. Utilized are recurrent statistical surrogate models that generate hourly 24-hour forecasts. To train these models, we use a selection of five sensor features collected over three years. Introduced is a multi-location modeling scheme that uniquely combines sensor data from nearby buoys as a novel methodology. This approach allows for more stable and accurate predictions compared to forecasting with single buoy data alone. Our experiments reveal that grouping six buoys yields the best forecasting performance. Furthermore, we improve model accuracy by integrating buoy data with numerical ocean models and applying a physics-regularized loss function. This technique mitigates the impact of missing or erratic data, leading to more dependable 24-hour forecasts. Our findings demonstrate that the combination of multiple-location modeling and physics-based regularization enhances the stability and accuracy of oceanic data forecasting.**

Keywords-Buoy Forecast; Multiple Location Forecast; Physics-Regularized; Numerical Models; Surrogate Models.

I. Introduction

Accurate forecasting of ocean and climate parameters is useful in industry and research. Climate analysis, ocean pollution management, extreme weather event tracking, and marine life monitoring, as examples, benefit from ocean modeling techniques [1][2]. Numerical models use initial data collected by observation sensors through buoys, ships, or satellites as inputs to their underlying physical equations. The result is a full coverage analysis of the physical features used to describe ocean and climate states. The European Centre for Medium-Range Weather Forecasts (ECMWF) research institute provides such forecast models for use in decision-making and analysis problems. These models rely on accurate observations of physical phenomena in two major ways. The first is the initialization values used as initial conditions for the physical equations. Then, after producing an analysis based on those initial conditions, a historical re-analysis of the model is generated by integrating the results with real-world sensor data through Data Assimilation (DA) techniques [3]. ECMWF's fifth reanalysis experiment (ERA5) dataset is a popular example of this and often used in statistical surrogate modeling tasks [4]. One source of observations is the fleet of free-floating ocean buoys anchored to fixed locations which are maintained by the National Data Buoy Center (NDBC). The processes of initialization and DA both require reliable, consistent, and high-quality observed sensor measurements to maintain accurate representations. Technical limitations or poorly calibrated sensors can yield noisy interpretations, and physical damage or scheduled maintenance can completely halt data collection in that location. Ocean and climate analysis or reanalysis, which rely on the steady stream of ocean sensor data provided by the NDBC, might benefit from short to midterm regional and sparse forecasts in this situation. This gives justification to investigate deep learning surrogate models to conduct sparse observation forecasting for data assimilation and other uses.

The sensor data is geographically sparse in the sense that there are collections of buoys within the same region which are separated by up to hundreds of kilometers. A surrogate model might be trained for individual analysis using single buoy data or multiple buoy data [5][6]. In both cases, only a single buoy is modeled at a time, ignoring any spatial complexities between surrounding buoys. Therefore, it is reasonable to investigate methodologies for modeling buoys at multiple locations within a shared context. One method for implementing spatial and temporal frequencies in a deep learning model is by introducing a graph neural network that leverages graph convolutions based on a buoy's spatial relations [7]. However, this approach relies on additional training and space overhead that is undesirable in a lightweight framework. Instead, we focus on less specialized deep learning frameworks. To apply a higher degree of spatial awareness in the model, all buoys in

a region of interest are included simultaneously. By expanding the radius to include more buoys, it can be demonstrated whether the machine learning surrogate benefits from the collective information. In addition, the impact on a specific buoy is investigated to determine if contextual clues from surrounding buoys aid in forecasting current conditions at the chosen location.

Surrogate modeling of ocean observations is useful in scenarios where fast and relatively accurate forecasts are needed. One problem with these models is that forecasts are accurate on short leads but lose accuracy as the time horizon increases [8]. For this reason, research into ways to combine the surrogate model with data assimilation, physical equations, or other numerical models have been investigated. The main goal is to reduce surrogate model error by incorporating physics understanding through alternative data sources. For example, machine learning models have been successfully incorporated with DA techniques for improved results [9][10]. Similarly, the concept of the Physics-Informed Neural Network (PINN) uses physical Partial Differential Equations (PDEs) solved in the loss function at training time to improve physical understanding [11]. This is similar to our methodology because both simulations and observations are used in the loss function. In contrast, our work seeks to find the best combinations of the contributions of precalculated models and observations. PINNs notably try to optimize the loss from observations to fit a particular model. The paradigm explored in this work is the combination of full-coverage numerical data and sparsely collected sensor data to produce a more stable model. To this end, NDBC data and ERA5 reanalysis data are combined to train surrogate models to forecast buoy-derived geographically sparse ocean observations.

Noisy data derived from sensors has a significant impact on training forecast models. Missing data is a compounding concern when analyzing individually collected sensor observations. To combat this problem, a methodology to combine sensor derived data with numerically modeled data at near-point locations was proposed in [12]. The surrogate model training procedure uses both numerical data and sensor derived outputs when calculating the loss score. These values are combined using a ratio of the two error scores and then back propagated through the architecture. The previous work is extended by using a modified forecasting methodology and data representation to see if similar increases in model accuracy are achieved. Also, numerically modeled features are removed from the model input to reduce numerical reliance post-training. In this way, the underlying physical calculations of the numerical models regularize the statistical surrogate models at training time alone.

The main considerations of this research are in improving techniques of machine learning with ocean data for the predictive modeling of observed phenomena. Specifically, the goal is to investigate whether a geographically sparse set of data can be structured in such a way that the predictive ability of the model is stabilized over 24-hour forecast cycles. The viability of combining sensor data with numerically derived data at training time is consequently explored to verify if further improvements to the surrogate models can be made. Combining data in this way is a new technique which has only been evaluated in two experimental situations [12][13]. This raises the question of whether the combined approaches are better used in unison or separately. So, our main contributions are listed as follows:

- We describe a novel training scheme that uses multiple buoys and their observed parameters as input into the model;
- Using the physics-regularized loss function, we use a grid search to find the best ratio of data for each combined feature;
- By identifying methodologies to improve surrogate model performance, we give further justification to the use of statistical models in an oceanographic context.

The paper begins with a summary of existing related work in Section II. Section III introduces the specific methodology used. The data sources, chosen data representation, physics-regularized loss function, and the selected architectures for the neural networks are detailed. The section concludes with the setup of the executed experiments. In Section IV, the experimental results are presented. The paper concludes with a summary and suggestions for future work in Section V.

II. Related Work

Research into forecasting ocean parameters using machine learning methodologies is abundant in literature. Reviewing recent innovations in modeling the physical parameters yield methodologies for wind modeling [14], ocean wave height/direction [6][15][16], air temperature [17], and sea surface water temperature [18][19]. Popular models for surrogate ocean forecasting include the Long-Short Term Memory (LSTM) model and the attention head transformer model [20][21]. Their use is highlighted in recent ocean parameter forecasting, so the use of these layers are adopted in this research [6][12][22]–[26].

The availability and quantity of ERA5 data make it a suitable choice for machine learning-based surrogate models. In recent years, surrogate models trained or otherwise supported specifically by ERA5 data are studied for use for regional wave modeling [15], weather forecasting [27][28], earth surface temperature modeling [29], and sea surface temperature forecasting [30][18]. The data is also used when enhancing sensor predictions, for example, in the case of satellite sensor models [31][32]. Notably, the focus is usually on one oceanic feature or phenomenon.

To better analyze real-world conditions at a specific location, ocean buoys are modeled. Traditionally, this can be done by training a statistical model on the data collected from a single location for prediction [5][16]. Otherwise, multiple observation locations are combined into a single time series to train and predict an individual location. This incorporates a sense of spatial awareness into the model [14]. With the advent of deep learning, dozens of locations are used to train a single model for a more generalized approach. The model is

trained and used to forecast multiple single-location buoys, for example [6][12]. Specialized architectures, like graph neural networks, are also used to increase spatial awareness [30].

Combining both physically collected and numerically modeled data together in an ML context is a wide-reaching methodology. PINNs combine PDE solutions with collected data at training time [11]. This approach is usually used to solve numerical equations directly [33]. Forecasting physical parameters such as sea surface temperature [34] or physical phenomena such as storm surges [35] are also explored. DA for machine learning models is also becoming a popular topic to improve surrogate results [10][28]. Most similar to our research is the use of both modeled and observed data in the training and/or inference of ocean parameters, as seen in [12]. That work differentiates itself from typical DA by combining the data during the training scheme itself, instead of as a reanalysis step. It is different from typical PINN models through the lack of modeling of differential equations in the training scheme, instead using pre-calculated numerical outputs.

The methodologies proposed in this research are novel with respect to surrogate ocean parameter forecasting because the highlighted works typically combine fewer ocean features for forecasting, while we use all buoy sensors available and combine with ERA5 data. This work is novel in the context of buoy forecasting because a Multiple-Location Model (MLM), as defined in this work, either has not been investigated or is difficult to find in literature. Lastly, this extends the previous work of [12] by examining the proposed loss function in a different context, modeling multiple buoys with a different feature set, instead of single buoys. A significant difference is the use of numerically modeled features only in the training dataset and not as an input into the surrogate at inference time. This better aligns the methodology as a class of DA that combines data observations and models at training time to improve model performance when initial conditions are poor at inference time.

III. Methodology

Introduced are methodologies to verify the following research goals. One aim is to see if it is possible to improve the forecasting result for the collection of selected buoy observations. The MLM technique is expected to improve accuracy when increasing the number of modeled buoys. If so, this will give insights on whether the model weights can internalize connection in spatially sparse regions of interest. The physics-regularized loss function is improved by using modeled features in the training scheme while removing them from the inference input. Thus, it must be investigated if the method still improves model accuracy and whether it can be combined with the new sparse modeling scheme.

A. Sources of Data

The National Data Buoy Center (NDBC) is a United States-based organization and a division of the National Oceanic and Atmospheric Administration (NOAA). NDBC is responsible for collecting real-world observations of a wide range of ocean and climate feature measurements. In total, 1311 stations are deployed with 243 owned and maintained directly by NDBC [36], and they have been collecting the observation data since 1970. Within the framework of this project's experiments, the utilization of standard meteorological historical data (STDMET) is employed. As the data is of real-world origin, not every station provides feature data for every year. Furthermore, the available data may contain missing values [37]. Sensors in real-world settings may experience malfunctions or localized noise that are not helpful in a model training context.

Table 1. Considered NDBC Features and ERA5 Equivalents

Measurement	NDBC Feature	ERA5 Equivalent
Air Temperature (°F)	ATMP	2m temperature
Air Pressure (hPa)	PRES	Mean sea level pressure
Dewpoint Temperature (°F)	DEWP	2m dewpoint temperature
Water Temperature (°F)	WTMP	Sea surface temperature
Wind Speed (kts)	WSPD	10m v-component of wind, 10m u-component of wind

The ERA5 dataset, generated by the ECMWF, represents the latest advancement in a series of global atmospheric reanalysis datasets. Reanalysis entails the assimilation of observational data and model simulations to generate a coherent and extensive dataset. Data assimilation leverages observations from ships, buoys, and satellites with physical laws to ensure historical data is modeled as accurately as possible. Within this project, the ERA5 dataset, specifically identified as "ERA5 hourly data on single levels from 1940 to present", is utilized [4]. Hourly estimates are offered for a diverse set of atmospheric and ocean circulation variables. This dataset consists of a regridded subset extracted from the full ERA5 dataset while preserving its native resolution.

NDBC standard meteorological dataset provides 14 measurements [37] while ERA5 provides 262 variables [4]. Five selected NDBC measurements are also modeled within the ERA5 dataset, and therefore, are collected for further analysis and use within the physics-regularized deep learning task. Combining the five NDBC measurements with the corresponding five ERA5 values yields the collection of data required for the upcoming physics-regularized loss function. These five selected features are the focus of the surrogate forecasting task. The selection collected from NDBC and the ERA5 counterparts are depicted in Table 1. We use approximately three full years of hourly data, from 2020 to 2022.

B. Multi-Location Modeling

Oceans are complex and elaborate systems with interconnected dynamics. Therefore, it is hypothesized that enriching training data with measurements from several locations will empower the neural network to recognize spatial dependencies

between nearby buoys. The spatial-temporal data from each location is integrated in a unified dataset following a novel design in which each instance encapsulates measurements of multiple stations at a particular timestamp. That is, at N buoy locations, the features in Table 1 are extracted and combined into a single vector of one-hour increments to create an MLM dataset. Corresponding locations for the buoys in the ERA5 data are combined in the same way. The ERA5 data is reserved separately for use in the specialized loss function described in Subsection C. The example in Figure 1 shows the partitioning process. Any number of buoys may be selected in this generalized approach.

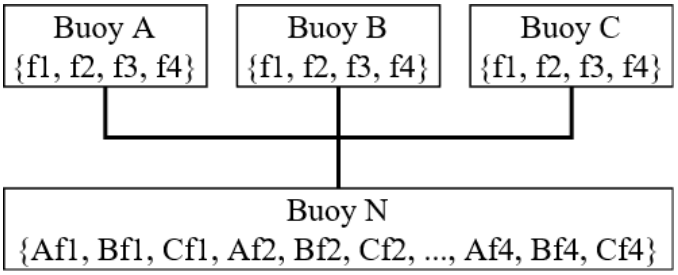

Figure 1. Example of the construction and final representation of the Multi-Location Modeling scheme.

In the figure, buoys A, B and C represent unique locations with a fixed buoy. Features [f1-f4] represent time series vectors of length T. Features are combined in alphabetical order into buoy N, such that sub-sampling a single element t gives the current condition of each buoy. The vector size of buoy N(t) is the number of features multiplied by number of buoys. The geographical locations influence one another implicitly through the neural network's hidden layers. Model weights take advantage of correlations learned at training time. This is contrasted against grid-based numerical models where forward regression is explicitly calculated with respect to adjacent grid cells.

Once the dataset has been generated, preprocessing is completed to make the problem more suitable for the deep learning architecture. First, we take the difference of each feature from t to $t+1$. This simplifies the forecast problem by reducing it to a gradient prediction problem. For each feature, the model only needs to produce an inference representing the rate of change. An inverse integration stage sums the recurrent predictions with the initial condition to get the real-world result. The data is normalized based on the mean and standard deviation seen in the training data alone, and the inverse is completed when investigating the results. Therefore, feature forecasts analyzed in the results section are in their respective scales.

C. Physics-Regularized Loss Function

The loss function of our deep learning models compares the training inference of the MLM vector with an individual observed value and numerically modeled value. By comparing the model inference to the NDBC and ERA5 data, two error scores are generated from the same prediction. Depending on interpolation or noise in the underlying data, one source will better represent the true conditions. So, a ratio of the two error scores is taken, and this error is used for back propagation in the model. The method is first proposed in [12]. The result is that, based on the ratio, the model is trained to approximate either source more strongly.

The disagreement between the datasets is proposed to improve the training procedure in two significant ways. The first is that the disagreement prevents overfitting to either individual source of training data. The second is that interpolated or distorted values are less likely to be frequent in both sources simultaneously. It is assumed that at least one source of data reasonably represents the underlying conditions for a given time step. Ultimately, the multi-location vector of buoy feature predictions is given along with the numerically derived values. Given the NDBC predictions and ERA5 values to be combined in the training procedure, the loss function is defined as follows in (1)-(6).

$$\Delta_1 = |\hat{y}_{\text{NDBC}} - y_{\text{NDBC}}| \tag{1}$$

$$\Delta_2 = |\hat{y}_{\text{NDBC}} - y_{\text{ERA5}}| \tag{2}$$

$$\Omega_{\text{coupled loss}} = (\alpha * \Delta_1) + ((1 - \alpha) * \Delta_2). \tag{3}$$

In (1) and (2), $\hat{y}$ represents the output vector of the surrogate model, while y represents the training ground truth vector. The source of modeling truth is determined by the subscript as y_{obs} or y_{model}. The error for each feature is weighted by coupling term α and represents a mixture of error calculated against two sources, ERA5 and NDBC. Importantly, α is constrained such that $0.0 \leq \alpha \leq 1.0$. When $\alpha = 0.0$, the NDBC term is completely shut off, and the model is only trained by comparing the ERA5 estimations. Otherwise, when $\alpha = 1.0$, only the NDBC data is used in training the model. Additional non-coupled features may be included in the surrogate and are defined as,

$$\Omega_{\text{model loss}} = |\hat{y}_{\text{ERA5}} - y_{\text{ERA5}}| \tag{4}$$

$$\Omega_{\text{observation loss}} = |\hat{y}_{\text{NDBC}} - y_{\text{NDBC}}|. \tag{5}$$

The remaining uncoupled features, as seen in (4) and (5), are used to collect error by comparing the predicted value with the relevant ground truth value. Each piecewise value is summed into the final loss function (6),

$$\Omega_{\text{total loss}} = \Omega_{\text{coupled loss}} + \Omega_{\text{model loss}} + \Omega_{\text{observation loss}}. \tag{6}$$

In this work, only the coupled loss is used. There are no uncoupled model or observation features, so $\Omega_{\text{total loss}} = \Omega_{\text{coupled loss}}$.

D. Deep Learning Architecture

Two deep learning architectures are considered for the forecasting task, given the prior defined multi-location dataset and the physics-coupled loss function. A LSTM unit model and an attention head transformer model are used. The two

models are chosen for their complexity and ability to generalize complex and recurrent time series data [38]. LSTM and attention head layers add more parameters to the hidden layers of the relatively shallow networks. The memory unit in the LSTM model particularly excels at storing prior knowledge for improved forecasts. The attention mechanism in the Transformer model is known for statistically weighing the importance of the previous input. The Transformer architecture typically implements more weights in the hidden layers of the model, so comparing this with the smaller LSTM model is often insightful. Both model architectures are implemented using the Python programming language and the TensorFlow machine learning platform [39].

TABLE 2. LSTM MODEL ARCHITECTURE BY LAYER FOR STAGE 2. THE TOTAL NUMBER OF TRAINABLE PARAMETERS IS 146,878.

Layer Type	Shape	Parameters
Input	*Variable*	0
LSTM	(N, 1, 128)	81,408
Dropout	N, 1, 128	0
LSTM	(N, 1, 64)	49,408
Dropout	N, 1, 64	0
LSTM	(N, 1, 32)	12,416
Dropout	N, 1, 32	0
LSTM	(N, 16)	3,136
Dense	(N, *Variable*)	510

TABLE 3. TRANSFORMER MODEL ARCHITECTURE BY LAYER FOR STAGE 2. THE TOTAL NUMBER OF TRAINABLE PARAMETERS IS 602,526.

Layer Type	Shape	Parameters
Input	*Variable*	3,968
Transformer Block	(N, 1, 128)	297,344
Transformer Block	(N, 1, 64)	297,344
Flatten	(N, 128)	0
Dense	(N, *Variable*)	3,870

Both models include dropout layers to prevent model overfitting during the training process. The dropout parameter is set to 0.1. The exact implementation of the LSTM model is found in Table 2 and the transformer model is in Table 3. In each table, N represents the variable batch size when training the model. The *Variable* input and output is equal to the product of the number of features and the number of buoys in each experiment. Therefore, each model's input and output shape is dependent on the number of ocean features and the number of buoys, as described in Figure 1. The Transformer blocks consist of multi head attention layers with four attention heads and feed forward dense layers of size 128 and dropout layers. Each model is trained for 100 epochs using a batch size of 64. The training procedure is set such that the model is given conditions at t and produces a forecast at $t+1$. By setting the input and output shape to the same length, the resulting model can be used to generate recurrent forecasts for any number of consecutive one-hour periods.

E. Experimental Test Case

Experiments using differing numbers of model locations and α ratio values are examined to verify the proposed methodology. The forecast results are centered on a particular buoy with the identification value of 42002. Buoy 42002 is located in the in the Gulf of Mexico. This buoy was chosen because of its central location relative to other buoys and because it has comparatively fewer missing values. Regardless of the number of buoys modeled, focusing on only one buoy assesses the hypothesis that the surrogate is improved when given spatial context. Considering the forecast results of all buoys is impactful, but outside the scope of this investigation.

When increasing the number of modeled buoys, results that are similar or better to the single-buoy model validate the proposed methodology for producing regional inferences. Increasingly better results indicate that the MLM approach either provides spatial context or model regularization through additional data. Tuning the α ratio is proposed to improve model forecast by combining observed and numerical data. So, an α value which produces better results than when $\alpha = 1.0$ (no regularization) is searched for.

For validation, the forecast ability of the surrogate models is examined for 24-hour intervals over the test data. The test data is comprised of the final three months of 2022. A recursive forecast scheme that uses the output of the proceeding inference as the input for the next inference is employed. The temporal resolution of the data is in one-hour increments, so 24 hourly forecasts are made for each interval. The model only uses the real conditions when making the first prediction in a cycle and when the test results are analyzed.

The Root Mean Square Error (RMSE) score (7) is used to determine the accuracy of the 24 hour forecast compared to the actual conditions at the buoy. The RMSE can be examined on a per-time step basis or as an average of all predictions.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

The MLM methodology is tested by increasing the number of buoys to forecast features in Table 1. Buoy 42002 is used to analyze ATMP, PRES, and WSPD, while buoy 42020 is analyzed for DEWP and WTMP. Two separate single-buoy experiments are conducted because no examined buoy contained all five features without extensive missing values. Using the coordinates of the central station, buoy 42002, and the Great Circle Distance [40] method, the distance to the remaining stations in the area was calculated. To evaluate the hypothesis, three datasets were created: one that exclusively covers a central station (Stage 0), one that covers stations within 600 km (Stage 1), and one that covers stations within 900 km (Stage 2). The exact specifications of the three experiments is found in Table 4. The geographical location of each buoy is displayed in Figure 2. To get baseline results, $\alpha = 1.0$ is selected, which means the entire training signal comes from the NDBC buoy dataset and the ERA5 data has no influence on the NDBC predictions.

TABLE 4. INCLUSION OF BUOYS PER EXPERIMENT. GENERAL LOCATIONS OF EACH BUOY ARE FOUND IN FIGURE 2. STAGE 0 USES TWO SEPARATE INDIVIDUAL BUOYS TO CAPTURE FEATURES MISSING IN EITHER EXPERIMENT.

Stage	Distance (m)	# of Buoys	Buoy List
Stage 0	-	1	42002; 42020
Stage 1	600	4	42002, 42019, 42020, 42035
Stage 2	900	6	42002, 42003, 42019, 42020, 42035, 42040

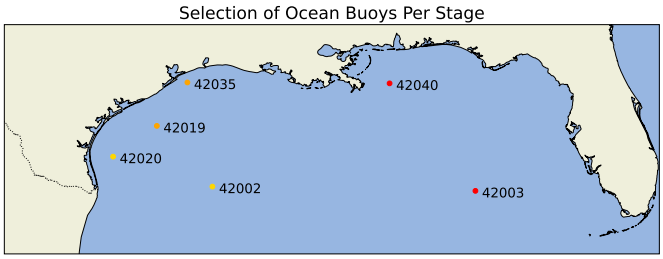

Figure 2. The geographical location of each buoy used in experiments. Stage 0 is labeled yellow, Stage 1 is labeled orange, and Stage 2 is labeled red.

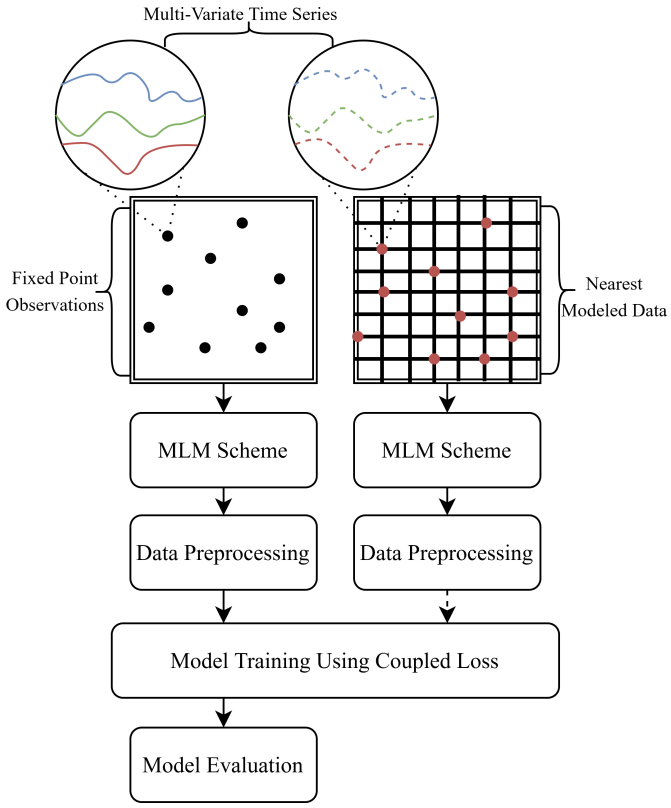

Figure 3. The data collection and processing pipeline.

Once the best number of buoys has been identified, the models re-examined with the physics-regularized loss. Finding the value for hyperparameter α that minimizes the error score is the main goal. A grid search of $\alpha \in [0.0, 1.0]$ with a step size of 0.05 is conducted. The surrogate model is retrained using the same random seed and for each α value. The minimal RMSE score on the test dataset denotes the best performing experimental setup.

For this experiment, the data collection and processing pipeline is divided into four stages, as outlined in Figure 3. First, the time series data is downloaded for each buoy from NOAA. Similarly, the ERA5 data is retrieved after selecting the appropriate geographical region and corresponding time periods. Second, the buoy's latitude, longitude, and time values are used to match the buoy data with the ERA5 data. Since these values may not align perfectly, we select the closest possible location and time points. Third, the MLM scheme is applied to format the forecast features as described in Figure 1. The same process is followed for the ERA5 data, which is stored in a separate vector for use in the coupled loss function (3). Finally, time series data processing is completed for all features. Missing data is interpolated, differencing is applied from t to $t+1$, and each feature is normalized based on the training data.

IV. RESULTS

The results for the MLM data technique are found for the LTSM model in Table 5 and the transformer model in Table 6. Each table displays the model, the examined buoy, and the calculated RMSE rounded to three decimal places for each stage. The average performance of all features is taken to find the overall best performing stage. From the initial stage experiments, a slight improvement is demonstrated for all features when comparing Stage 0 to Stage 1 and/or Stage 2. This implies that the additional data used in the MLM data scheme has an impact on model stability.

One outlier in the experiments is found in water temperature (WTMP), which yields almost the same result in each experiment. We propose the behavior is due to two main factors. The first is that most buoys are missing WTMP data, a differenced forecast of near zero is preferred by the model. The second is that the ERA5 WTMP data, used when $\alpha < 1.0$, is only updated every 24-hours. Therefore, the regularizing data is also biased in the same way. This important lesson shows that poor data, when collected for both sources, produces an uninformed model. Neither the MLM scheme nor the physics-regularized loss function can improve results in this case.

Model improvements when adding additional buoy data is shown in Figure 4. In the figure, consecutive 24H forecasts are generated for the air temperature (ATMP) feature for all three stages. In Stage 0, the prediction is unstable for long

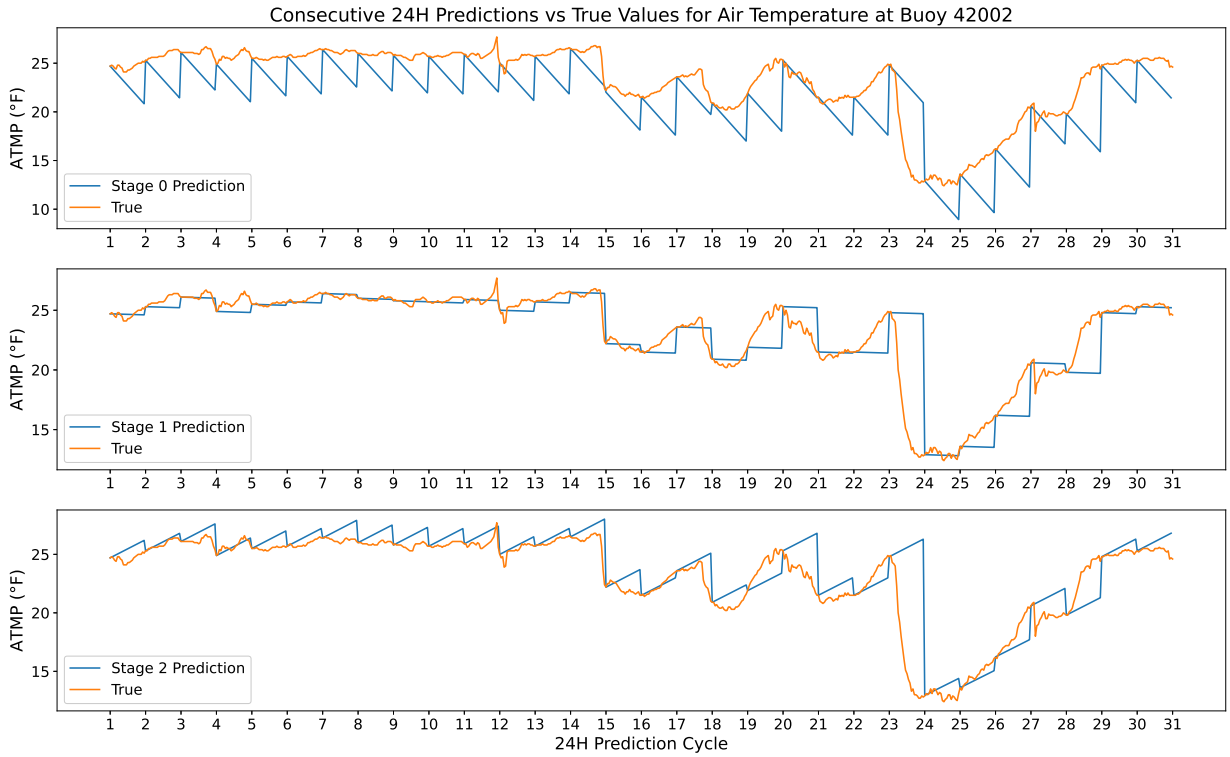

Figure 4. Improvements in Transformer model stability when adding additional buoy data to the prediction vector. Increasing buoy count results in more conservative predictive behavior.

TABLE 5. RESULTS OF CONSECUTIVE 24H FORECASTS USING THE LSTM MODEL FOR EACH STAGE. ATMP, PRES, AND WSPD ARE TESTED USING BUOY 42002 WHILE DEWP AND WTMP ARE COLLECTED FROM BUOY 42020.

LSTM	Buoy	Stage 0	Stage 1	Stage 2
ATMP	42002	1.909	1.736	1.733
PRES	42002	12.683	12.483	12.132
WSPD	42002	2.698	2.513	2.516
DEWP	42020	3.218	3.447	3.164
WTMP	42020	0.625	0.625	0.625
Average		4.227	4.161	**4.034**

TABLE 6. RESULTS OF CONSECUTIVE 24H FORECASTS USING THE TRANSFORMER MODEL FOR EACH STAGE. ATMP, PRES, AND WSPD ARE TESTED USING BUOY 42002 WHILE DEWP AND WTMP ARE COLLECTED FROM BUOY 42020.

Transformer	Buoy	Stage 0	Stage 1	Stage 2
ATMP	42002	2.851	1.733	1.940
PRES	42002	12.677	12.460	12.302
WSPD	42002	2.560	2.595	2.514
DEWP	42020	6.262	3.158	3.160
WTMP	42020	0.625	0.625	0.625
Average		4.995	4.114	**4.108**

forecast horizons. Applying more data via the MLM scheme regularizes the behavior to a more stable outcome. Stage 1 shows results that are very stable with nearly no change. This performs very well in low-change periods. In Stage 2, the addition of more data to forecast yields a less stabilized result overall. The stability is better than Stage 0, and the forecast might be more useful in drastically changing systems.

Following the stage experiments, various α values are iterated over to validate whether the physics-regularized loss function improves the MLM data scheme. The experiment is conducted for both model architectures and all stages. Results when $\alpha = 1.0$ are equivalent to the MLM stage experiments. The average results of these tests can be found in Table 7. In the table, all but one model and stage combination yield

TABLE 7. AVERAGE α RESULTS FOR ALL MODELS AND STAGES. THE MOST PERFORMANT EXPERIMENTS ARE SET IN BOLD.

Model: Stage	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
LSTM: 0	5.072	5.037	6.055	4.800	4.647	4.899	5.419	5.063	5.152	**3.956**	4.320
LSTM: 1	5.174	4.921	4.708	4.575	4.652	5.304	8.051	6.33	5.526	4.415	**4.028**
LSTM: 2	4.647	4.412	**3.999**	4.74	5.079	4.66	4.962	6.126	4.473	4.432	5.218
Transformer: 0	5.328	5.187	5.361	5.631	5.761	5.252	5.0140	5.088	4.9780	5.168	4.731
Transformer: 1	5.696	5.604	5.433	5.911	5.273	6.129	5.061	4.964	5.338	5.165	4.938
Transformer: 2	5.066	5.119	5.160	5.251	4.929	4.960	4.773	4.673	4.838	4.597	4.853

Model: Stage	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
LSTM: 0	5.242	6.050	4.670	4.420	4.378	4.576	5.138	4.634	4.818	4.227
LSTM: 1	5.381	6.475	4.688	4.186	4.492	4.971	4.905	4.986	5.257	4.161
LSTM: 2	4.420	4.455	5.522	4.084	4.596	4.878	4.692	4.603	4.244	4.034
Transformer: 0	4.844	4.511	4.543	4.558	4.604	4.317	4.597	**4.088**	4.371	4.995
Transformer: 1	5.210	4.824	4.723	5.008	4.928	4.884	4.605	4.48	4.205	**4.114**
Transformer: 2	4.370	**4.106**	4.495	4.505	4.374	4.364	4.604	4.178	4.416	4.108

TABLE 8. RESULTS OF ITERATIVE α RATIO TESTING USING THE LSTM MODEL AT STAGE 0. WHEN $\alpha = 0.0$, THE MODEL IS TRAINED USING EXCLUSIVELY ERA5 DERIVED DATA. WHEN $\alpha = 1.0$, THE MODEL IS TRAINED USING EXCLUSIVELY NDBC BUOY DATA. FOR $0.0 < \alpha < 1.0$, A MIXED FORMULATION OF (3) IS USED.

Feature	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ATMP	1.772	1.741	1.979	1.779	1.75	1.81	1.745	1.746	1.863	1.732	1.957
PRES	16.883	15.982	9.825	14.376	13.87	14.713	14.463	14.406	13.879	11.021	12.461
WSPD	2.541	3.311	14.571	3.466	3.453	2.689	2.543	3.061	5.158	2.784	3.355
DEWP	3.539	3.525	3.277	3.755	3.535	4.659	7.720	5.479	4.234	3.617	3.200
WTMP	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625

Feature	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
ATMP	1.731	1.757	2.027	2.319	1.752	1.789	1.807	2.441	2.127	1.909
PRES	13.09	13.33	13.670	12.738	12.212	11.694	12.877	12.738	13.072	12.683
WSPD	6.168	6.31	3.203	3.007	2.977	3.204	5.803	2.571	2.483	2.698
DEWP	4.597	8.223	3.820	3.413	4.325	5.568	4.577	4.794	5.782	3.823
WTMP	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625

results that are superior to the $\alpha = 1.0$ (all NBDC data) model. However, typically the most significant reduction is found for Stage 0. This implies that the stability gained from the MLM modeling is significant enough that the additional regularization gained from the coupled loss function is minimal. In other words, when there is less available geographical context in the data representation, numerical model regularization is more impactful. The results from the LSTM Stage 0 experiment are highlighted in Table 8. We highlight that the combination of data may not be viable for all features at once. That is, the feature yields minimal results at various α values. This suggests that multiple α ratios may be used, one for each feature. This idea is explored further in similar ongoing research [13] but is outside the scope of this work.

To better understand how the error is reduced among the α experiments, the absolute error generated from $\alpha = 1.0$ and $\alpha = 0.9$ for the Transformer Stage 0 experiment is compared in Figure 5. The bottom figure shows the difference between the two errors. When the value is greater than zero, this represents places where $\alpha = 0.9$ is more performant than the original $\alpha = 1.0$ model. In general, the coupled loss function regularizes the model in the same way as the MLM data scheme. Less exaggerated forecasts keep the model stable over longer horizons. In cases not highlighted in figures, it was observed that $\alpha < 1.0$ produced models that were better aligned to general ocean conditions by proactively forecasting changes in the environment. The final RMSE of these forecasts were worse on average but showed the influence of the ERA5 dataset on the models. Considering the results of both experiments, we have improved the forecast error and achieved roughly the best performance possible using the model architecture and data set. The improvements to model performance are capped because of the size of the models and the amount of training data used. Three years of buoy data at only six locations is not enough data to support a statistical model with millions of parameters, hence the smaller model architectures used in this experiment. Increasing the number of model parameters and amount of training data should result in a surrogate model that is more robust to real-world conditions, allowing the RMSE to be reduced further. These changes would provide diminishing returns once the model is large enough. Additional feature engineering for more informative features can also improve baseline results.

When comparing these baseline results to those generally

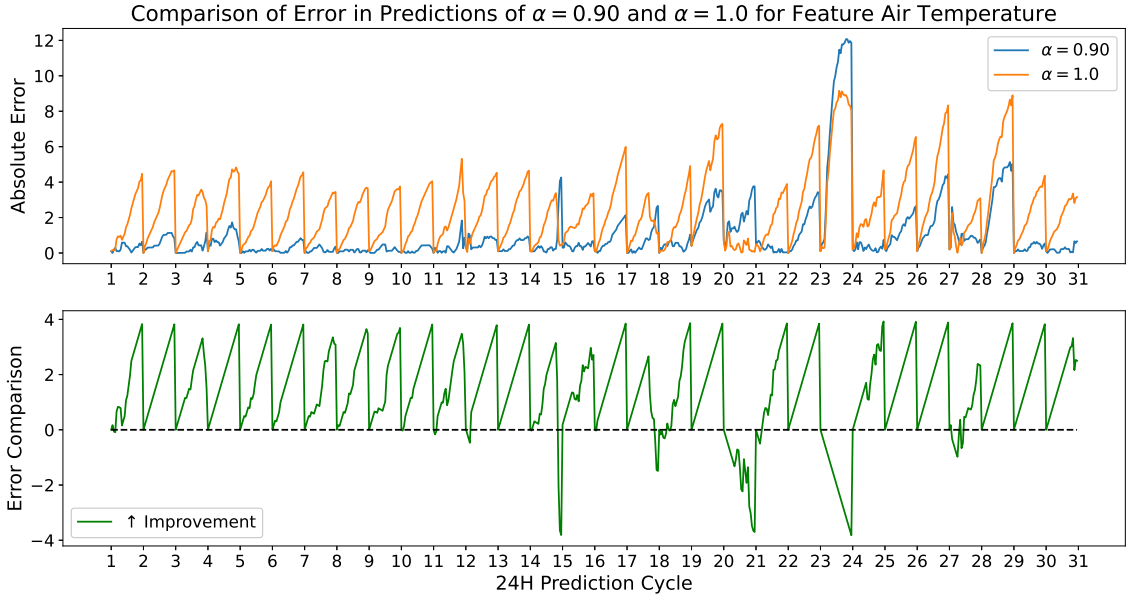

Figure 5. Difference in absolute error of the Transformer Stage 0 forecasts when comparing $\alpha = 0.9$ and $\alpha = 1.0$ (no regularization) for ATMP. The use of model data when training the observation model helps to prevent large swinging changes in the recursive forecast.

achieved in the most similar research to this [12], they are generally worse. This is explained by the smaller model size and dataset. However, the increased number of forecast steps is also a significant factor. The surrogate models in this experiment forecast three times as many periods, 24 instead of the previous eight, which means that general model stability is more important. Also, the previous work considers the combined result of over 100 buoys. Using a more similar experimental setup would likely continue to improve the results seen here. However, increasing the number of modeled buoys would continue to increase the inference vector size to a potentially cumbersome level. Since the input and output vectors of our models is the product of the number of buoys and number of features the required number of trainable parameters would increase considerably. A future investigation might select buckets of nearby buoys to model using the MLM scheme. This would increase the data pool without modeling hundreds of buoys in a single inference.

The MLM scheme improved model accuracy as the number of buoys increased. The distance between the buoys in Stage 1 and Stage 2 is large enough that local conditions are unlikely to affect one another. Therefore, it does not seem like a direct geographical influence is the determining factor of inferences. Instead, it seems most likely that utilizing more data in the input vector improves stability when training the weights of the neural network. To verify this hypothesis, further studies which use the MLM scheme in a variety of ways can be examined. First, using the MLM scheme on very distinct and far away locations should reveal if model behaviors are stabilized from the increased modeling space alone. Then, nearby observation points, which are geographically relevant to one another, can be added one-by-one until diminishing returns are found. This would support that the neural network is internalizing nearby behaviors. Although a more stable output is produced in the context of this research, other mechanisms for adding geospatial context should also be explored.

The physics-regularized loss function displayed the expected results of decreasing RMSE when the best α value is found. This compares to the results in [12], where a similar reduction in error through more stabilized results was seen. Contrary to those results, the error reduction seen is lesser in magnitude, most likely because the testing dataset is significantly larger in that work. Similarly to the previous work, some α values yielded results which were significantly worse than when no regularization is provided. This is typically the case when $\alpha = 0.0$ and most often because of the poor performing PRES feature. When $\alpha = 0.0$ this is equivalent to only using numerical model data to predict observation data. So, this behavior is most likely explained by misalignment between the observation ground truth and the numerically modeled PRES values from ERA5.

Importantly, the ERA5 data was removed from the inference input, and this is novel when comparing to the previous work. This shows the methodology can be used without the use of numerical models as an input parameter after the surrogate has been trained. This was an important contribution because it allows the methodology to be used more flexibly in real-world examples. Numerically modeled data is likely to be available when training a surrogate model, because this is almost always done using historical data. However, numerical models can

take hours to run for high-quality analysis. In situations where observation forecasts are needed immediately, this bottleneck would be extremely detrimental.

Finally, both techniques were separately found to regularize the results below the baseline. When combining both MLM and the physics-regularized loss, only a limited performance increase was found. This supports the conjecture that the model has reached a theoretical limit via the data selection and model architecture. Therefore, the use of both methodologies together should be reserved for those cases where more model regularization is required or when the problem is well-suited for both techniques. The MLM technique is recommended for those tasks where a static number of observation points need to be modeled. The physics-regularized loss can be implemented in any problem where multiple sources of data representing the same phenomenon are readily available.

V. Conclusion and Future Work

In this paper, a novel combination of a MLM scheme and a physics-regularized loss function was investigated for deep learning models. Fixed-location ocean buoy observations were used to validate this methodology in a real-world context. The buoy inference model was used to recurrently forecast 24-hour intervals over one month to validate results. Combining multiple buoy locations and the relevant features into a shared inference vector using the proposed MLM scheme improved model performance by stabilizing the inferences over longer time horizons. Multiple locations equaling four and six locations both yielded superior results over the single-buoy model. Improving the physics-regularized loss approach by removing numerical models from the surrogate input was also a success. The grid search for the best α value increased the performance the most in Stage 0 experiments. The combination of both MLM and coupled loss produces more accurate forecasts, but the magnitude of the improvement is lesser than when using either technique alone. This implied we reached the best results for this configuration of model architecture and data set. Although the experimentation was focused on a single buoy, the trends observed are expected to hold for the other forecasted buoys.

All together, it is proposed that the findings demonstrate enhanced stability and accuracy of oceanic data forecasting when using MLM and the physics-regularized loss. This is a practical surrogate for systems where multiple fixed-location observations (e.g., buoys) need to be forecasted simultaneously and in quick succession. Situations where a fleet of buoys have some missing values, due to buoy damage or scheduled maintenance, can benefit from this type of modeling. If only some data is missing, known values can be injected as a model input to support forecasts in regions where data is missing. These approximate observations can be used in place of sensor data while buoy maintenance is conducted. Finally, in this work the ERA5 numerical model data was used when training the model, but no features were used as part of the input during inference time. Practically, this technique allows for more flexibility in real world scenarios while still giving the model some physics-based regularization.

In future work, the same methodology should be compared against other data to verify if the combination of MLM and coupled loss would see more significant increases in model accuracy. Examinations considering all buoys in the forecast model should be conducted. Further comparisons to other geospatial context models, like graph neural networks, and their integration with the coupled loss function would also be extremely relevant. Explorations of the MLM structure and physics-regularized loss should also be adapted for use with other methods of combining physical knowledge or observations. Moreover, more complicated integration schemes like Runge-Kutta can be implemented to further improve performance/stability. Therefore, the use of the proposed methodology with PINNs or data assimilative machine learning is a promising potentiality in the right circumstances.

Acknowledgments

This work was partly supported by the U.S. Department of the Navy, Office of Naval Research (ONR), and Naval Research Laboratory under contracts N0073-16-2-C902 and N00173-20-2-C007, respectively. The work of Austin Schmidt was funded by a SMART (Science, Mathematics and Research for Transformation) Department of Defense (DoD) scholarship for service. The work of Elias Sandner was equally funded by the contract N00173-20-2-C007 and the Austrian Marshall Plan Foundation. The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] M. J. Kaiser and A. G. Pulsipher, "The impact of weather and ocean forecasting on hydrocarbon production and pollution management in the gulf of mexico," *Energy policy*, vol. 35, no. 2, pp. 966–983, 2007.

[2] A. J. Hobday, C. M. Spillman, J. Paige Eveson, and J. R. Hartog, "Seasonal forecasting for decision support in marine fisheries and aquaculture," *Fisheries Oceanography*, vol. 25, pp. 45–56, 2016.

[3] F. Bouttier and P. Courtier, "Data assimilation concepts and methods march 1999," *Meteorological training course lecture series. ECMWF*, vol. 718, p. 59, 2002.

[4] C3S CDS, *ERA5 hourly data on single levels from 1940 to present*, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), [retrieved: October, 2024], 2023. DOI: 10.24381/cds.adbb2d47.

[5] G. Ibarra-Berastegi *et al.*, "Wave energy forecasting at three coastal buoys in the bay of biscay," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 4, pp. 923–929, 2016.

[6] P. Pokhrel, E. Ioup, J. Simeonov, M. T. Hoque, and M. Abdelguerfi, "A transformer-based regression scheme for forecasting significant wave heights in oceans," *IEEE Journal of Oceanic Engineering*, vol. 47, no. 4, pp. 1010–1023, 2022. DOI: 10.1109/JOE.2022.3173454.

[7] Z. Wu *et al.*, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, Jan. 2021, ISSN: 2162-2388. DOI: 10.1109/tnnls.2020.2978386.

[8] G. Reikard and W. E. Rogers, "Forecasting ocean waves: Comparing a physics-based model with statistical models," *Coastal Engineering*, vol. 58, no. 5, pp. 409–416, 2011.

[9] M. Bocquet, "Surrogate modeling for the climate sciences dynamics with machine learning and data assimilation," *Frontiers in Applied Mathematics and Statistics*, vol. 9, p. 1 133 226, 2023.

[10] P. Pokhrel, M. Abdelguerfi, and E. Ioup, "A machine-learning and data assimilation forecasting framework for surface waves," *Quarterly Journal of the Royal Meteorological Society*, vol. 150, no. 759, pp. 958–975, 2024.

[11] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[12] A. B. Schmidt, P. Pokhrel, M. Abdelguerfi, E. Ioup, and D. Dobson, "Forecasting buoy observations using physics-informed neural networks," *IEEE Journal of Oceanic Engineering*, pp. 1–20, 2024. DOI: 10.1109/JOE.2024.3378408.

[13] A. B. Schmidt, P. Pokhrel, M. Abdelguerfi, E. Ioup, and D. Dobson, "An algorithm for modelling differential processes utilising a ratio-coupled loss," *TechRxiv*, 2024.

[14] Y.-Y. Hong, C. L. P. P. Rioflorido, and W. Zhang, "Hybrid deep learning and quantum-inspired neural network for day-ahead spatiotemporal wind speed forecasting," *Expert Systems with Applications*, vol. 241, p. 122 645, 2024.

[15] L. Huang, Y. Jing, H. Chen, L. Zhang, and Y. Liu, "A regional wind wave prediction surrogate model based on cnn deep learning network," *Applied Ocean Research*, vol. 126, p. 103 287, 2022.

[16] S. Londhe and V. Panchang, "One-day wave forecasts using buoy data and artificial neural networks," in *Proceedings of OCEANS 2005 MTS/IEEE*, IEEE, 2005, pp. 2119–2123.

[17] Q. F. Qian, X. J. Jia, and H. Lin, "Machine learning models for the seasonal forecast of winter surface air temperature in north america," *Earth and Space Science*, vol. 7, no. 8, e2020EA001140, 2020.

[18] X. Yu *et al.*, "A novel method for sea surface temperature prediction based on deep learning," *Mathematical Problems in Engineering*, vol. 2020, no. 1, p. 6 387 173, 2020.

[19] S. Wolff, F. O'Donncha, and B. Chen, "Statistical and machine learning ensemble modelling to forecast sea surface temperature," *Journal of Marine Systems*, vol. 208, p. 103 347, 2020.

[20] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132 306, 2020.

[21] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[22] K.-S. Kim, J.-B. Lee, M.-I. Roh, K.-M. Han, and G.-H. Lee, "Prediction of ocean weather based on denoising autoencoder and convolutional lstm," *Journal of Marine Science and Engineering*, vol. 8, no. 10, p. 805, 2020.

[23] Y. Liu *et al.*, "Spatiotemporal wave forecast with transformer-based network: A case study for the northwestern pacific ocean," *Ocean Modelling*, p. 102 323, 2024.

[24] Q. Zhang, H. Wang, J. Dong, G. Zhong, and X. Sun, "Prediction of sea surface temperature using long short-term memory," *IEEE geoscience and remote sensing letters*, vol. 14, no. 10, pp. 1745–1749, 2017.

[25] W. Chang *et al.*, "Real-time prediction of ocean observation data based on transformer model," in *Proceedings of the 2021 ACM International Conference on Intelligent Computing and its Emerging Applications*, 2021, pp. 83–88.

[26] X. Liu, T. Wilson, P.-N. Tan, and L. Luo, "Hierarchical lstm framework for long-term sea surface temperature forecasting," in *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2019, pp. 41–50.

[27] A. Chattopadhyay, M. Mustafa, P. Hassanzadeh, E. Bach, and K. Kashinath, "Towards physics-inspired data-driven weather forecasting: Integrating data assimilation with a deep spatial-transformer-based u-net in a case study with era5," *Geoscientific Model Development*, vol. 15, no. 5, pp. 2221–2237, 2022.

[28] M. Adrian, D. Sanz-Alonso, and R. Willett, "Data assimilation with machine learning surrogate models: A case study with fourcastnet," *arXiv preprint arXiv:2405.13180*, 2024.

[29] R. Niu *et al.*, "Multi-fidelity residual neural processes for scalable surrogate modeling," *arXiv preprint arXiv:2402.18846*, 2024.

[30] J. Kim, T. Kim, J.-G. Ryu, and J. Kim, "Spatiotemporal graph neural network for multivariate multi-step ahead time-series forecasting of sea temperature," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106 854, 2023.

[31] B. Kesavakumar, P. Shanmugam, and R. Venkatesan, "Enhanced sea surface salinity estimates using machine-learning algorithm with smap and high-resolution buoy data," *IEEE Access*, vol. 10, pp. 74 304–74 317, 2022.

[32] R. Zhang, Q. Liu, R. Hang, and G. Liu, "Predicting tropical cyclogenesis using a deep learning method from gridded satellite and era5 reanalysis data in the western north pacific basin," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2021.

[33] T. de Wolff, H. Carrillo, L. Marti, and N. Sanchez-Pi, "Towards optimally weighted physics-informed neural networks in ocean modelling," *arXiv preprint arXiv:2106.08747*, 2021.

[34] T. Yuan *et al.*, "A space-time partial differential equation based physics-guided neural network for sea surface temperature prediction," *Remote Sensing*, vol. 15, no. 14, p. 3498, 2023.

[35] C. Fu, J. Xiong, and F. Yu, "Storm surge forecasting based on physics-informed neural networks in the bohai sea," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 2718, 2024, p. 012 057.

[36] NDBC, *National Data Buoy Center*, https://www.ndbc.noaa.gov/, [retrieved: October, 2024].

[37] NDBC, *Measurement Descriptions and Units*, https://www.ndbc.noaa.gov/faq/measdes.shtml, [retrieved: October, 2024].

[38] M. Haghbin, A. Sharafati, D. Motta, N. Al-Ansari, and M. H. M. Noghani, "Applications of soft computing models for predicting sea surface temperature: A comprehensive review and assessment," *Progress in earth and planetary science*, vol. 8, no. 1, pp. 1–19, 2021.

[39] M. Abadi *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015.

[40] R. Bullock, "Great circle distances and bearings between two locations," *MDT, June*, vol. 5, pp. 1–3, 2007.