Physics-Regularized Buoy Forecasts: A Multi-Hyperparameter Approach Using Bounded Random Search

Austin B. 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

Md Meftahul Ferdaus
GulfSCEI
University of New Orleans
New Orleans, United States
email: mferdaus@uno.edu

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

***Abstract*—One challenge in oceanographic analysis is the need for accurate initial conditions collected from physical buoys. Temporary sensor outages or noisy conditions can hinder the data collection process. Machine learning surrogate models offer short-term coverage during outages. This study presents a methodology for regularizing machine learning models that predict buoy observations by utilizing multiple data sources. A previous work introduced a ratio-coupling hyperparameter to combine numerically modeled data and ocean observations when calculating training loss. However, applying one ratio across all features failed to capture the unique characteristics of different data sources. To overcome this limitation, this work investigates a multiple-hyperparameter loss function to independently manage the contribution of each data source per feature. A bounded random grid search explores the hyperparameter space to find ratios which produce superior results compared to the single-ratio approach. Surrogate models are validated at the same 88 fixed locations as the previous paper for a direct comparison. The experimental results suggest that this multi-ratio methodology can offer more reliable forecasts over a 24-hour period by applying the correct weight for each pairing of observed feature and numerical model source.**

Keywords-Surrogate; HYCOM; ERA5; Deep Learning; Buoy Forecasting.

I. INTRODUCTION

Modeling ocean and climate conditions is very important in industry settings and oceanographic research. Tasks like climate modeling, marine life population surveys, and tsunami monitoring, all rely on accurate understandings of ocean conditions [1][2][3]. Whether directly or indirectly, each of these tasks depends on the accurate initial conditions gathered from physical sensors. For that reason, this work focuses on machine learning modeling of sensor-derived data to produce short-term forecasts during temporary outages. The resulting forecasts can be used in place of observations for direct analysis, initial conditions for numerical models, or as data assimilation when performing a reanalysis. The types of conditions directly considered in this work include Sea Surface Temperature (SST), air pressure, and gust strength. Anomalies in SST can significantly affect accurate weather prediction [4]. Air pressure predictions are helpful for forecasting energy gain in photovoltaic systems [5] and intelligent weather forecasting systems [6]. Strong gusts cause severe damage in thunderstorms and are a forecasting target in machine learning tasks such as [7]. So, accurate predictions of these interconnected phenomena are highly relevant.

Whether considering sensor-derived observations or a carefully derived numerical solutions, there are often multiple ways to represent ground truth in a physical system. The numerical features used to describe our oceans and atmosphere are simply approximations of the underlying conditions. Systematic errors in data collection, physical errors in sensors, and spatiotemporal gaps in availability make observations unreliable by themselves [8][9]. Likewise, numerically modeled data have spatial and temporal discretization errors or miscalculations from strongly nonlinear interactions [10]. Imperfect approximations always exist, so combining various data sources becomes a worthwhile endeavor to reduce the inherent biases of each individual source. Traditionally, the use of data assimilation systems to improve models has seen great success. Reanalysis of numerical models with 4D variational data assimilation and Kalman filters improve historical model data to high accuracy [11][12]. This process yields high-quality training data for statistical surrogate models. However, data assimilation methods can only be used retroactively or when observation data is readily available. They also do not typically address errors in the underlying numerical model. From a machine learning context, multiple data sources can be combined as part of the training process instead. Due to multiple representations of truth, there is potentially more than one source of relevant training data. For example, SST can be represented by either a numerical model or by sensor observations. Therefore, improving forecasts by selecting the best source of truth for the training signal is a valuable goal.

To experiment with machine learning solutions for ocean

modeling problems, statistical surrogate models trained on a mixture of observed data and numerically modeled data are employed. A surrogate model is any model which is an approximation of a system without being numerically derived. This includes pure data driven approaches and also hybrid-physics approaches, like Physics-Informed Neural Networks (PINNs) [13]. Regardless of the method, surrogate model are trained to approximate generalized behaviors of the underlying system. To improve surrogate model performance, data is combined from fixed observation sensors and numerical models for a richer feature set. The combination can be formulated as a mixture of data assimilation and machine learning [14]. Conversely, the entire physical phenomenon can be modeled together by directly training a surrogate model with numerical outputs and sensor data [15][16]. This work follows the latter paradigm where the physical phenomenon is directly modeled. Specifically, the experimental design follows the problem domain presented in [16].

In [16], a specialized loss function was introduced which coupled noisy observation data and imperfect numerically modeled data. The buoys for observation of the ocean in fixed locations collect sensor data from around the coast of the United States, and surrogate models were used to forecast their observed features. To improve the regional surrogate model stability, historical ocean modeling data from the same regions are added to the training set. The Hybrid Circulation Ocean Model (HYCOM) [17] and the fifth reanalysis experiment of the European Center for Medium-Range Weather Forecasts (ERA5) [18] were chosen for their selection of global climate and weather features. Features that are available in both observed and modeled sources were coupled together in the loss function of the training procedure. A performant ratio of the loss signals was identified by balancing error between the surrogate inference and the two sources of training data per coupled feature. That is, each coupled feature has both an observed value and a corresponding numerically modeled value from either HYCOM or ERA5. A limitation in this methodology was the use of a single hyperparameter to control the ratio of all coupled features. To improve the identified limitation, the single hyperparameter is redefined as a vector of N hyperparameters. Consider that the ocean features are combined with multiple numerical sources. One numerical model may be well tuned to the underlying conditions of one feature and necessitate a stronger contribution to the training signal. If the other numerical solution does not align as well with the ground truth, an independent hyperparameter allows the training signal of that source to be reduced, while the other remains a major contributor in the loss calculation. Consequently, the main contributions of this paper are as follows:

- A surrogate training scheme is defined and validated that uses a physics-regularized loss function to independently combine two sources of data for the characteristics of the ocean K.
- Showing improvements in surrogate performance justifies the use of statistical models in oceanographic analysis.
- Finding improvements in combining two data sources for ocean features promotes continued exploration of data combination techniques during model training time.

The remainder of the paper is organized into the following sections. In Section II, the related work identifies similar research and contrasts them to this one. The main research goal is identified in relation to the previously identified work. In Section III the methodology is specified. The experimental dataset details are outlined, the improvements for the physics-regularized loss are detailed, and the deep learning architecture used is described. The experimental setup used for validation is provided for reproduction. Subsequently, Section IV is the Results section where the experimental findings and their impacts on the methodology are described. Finally, in Section V, the major contributions are reiterated and future considerations are identified.

II. Related Work

Buoy forecasting is investigated in some statistical learning contexts similar to this work. Models are trained using one buoy [19][20] or multiple buoys [21] for a region of interest. Most often, buoy observation forecasting focuses on the analysis of a single buoy, instead of many buoys in a variety of conditions. In one work, a collection of buoys are integrated into the input and output vector [22], but the rigid design of the architecture requires less flexible batch forecasting. Comparatively, there do exist works where a deep learning model is used for generalized buoy forecasting [15][16]. This research follows the scheme of generalized deep learning models to forecast a wide range of buoys, given their initial conditions.

The numerical models HYCOM and ERA5 are used in machine learning-based projects as training data for surrogate modeling tasks. In the case of HYCOM, the modeled data is used in machine learning forecasting tasks for ocean conditions [16] and sea surface salinity [23]. HYCOM data is also used to combine observations with modeled data in a machine learning context to parameterize typhoon-ocean interactions [24]. The ERA5 data is used more commonly, most likely due to its ease of access and high number of modeled parameters. It is used as training data for regional wave modeling [25], weather forecasting [26][27], earth surface temperature modeling [28], and sea surface temperature forecasting [29][30]. Numerically modeled data is also used when enhancing sensor predictions, for example, in the case of satellite sensor models [31][32]. Usually only one oceanic feature is forecasted, in contrast to this work. Also, when enhancing sensor forecasts with numerically modeled data, it is less common to combine more than one numerical model.

Recurrent Neural Networks (RNN) and attention models are classically used in time series-based modeling problems, making them a natural choice for oceanic forecasting. The Gated Recurrent Unit (GRU) is one type of RNN unit that employs an update and reset gate as part of the architecture for improved

temporal learning [33]. Oceanographic modeling that uses a GRU-based architecture has been used for ocean current prediction [34] and chlorophyll concentration forecasting [35]. One work similarly focuses on buoy sensor SST forecasting using GRU architectures [36]. However, the methodology differs from this work in significant ways including the model architecture, number of features forecasted, and the use of numerical models.

The physics-regularized loss function for training surrogate models has been examined in two experimental papers [16][22] and one theoretical analysis [37]. The methodology proposed in this work bridges the gap between two of those papers. In the original paper, it was proposed in the concluding section that separating the ratio-controlling hyperparameter would continue to improve results [16]. The research in this paper directly extends this previous work by testing that hypothesis using the same experimental setup. Although [37] uses a similar multiple-parameter scheme we propose here, the work does not highlight this fact. There is an assumption separating the hyperparameters is an improvement on the methodology, but no formal study was ever undertaken. Therefore, this is the first study using the physics-regularized loss function that investigates whether the use of multiple hyperparameters improves the physics-regularized model by directly comparing against the original implementation.

III. Methodology

The presence of multiple, potentially biased representations of truth within a domain presents a challenge for optimizing machine learning models. A loss function that effectively leverages these diverse sources of truth can minimize test error. One approach is to use a loss function that balances the contributions of different data sources using a coupling hyperparameter, $\lambda \in [0, 1]$, which determines the ratio of each source's influence [16]. The following subsections describe the methodologies used to extend the previous paper and answer the main research question. That is, whether splitting the single coupled hyperparameter into a vector of independent coupled hyperparameters improves the result. Specifically, this study investigates whether this proposed modification results in continued improvement under the same experimental conditions.

A. Dataset Details

Buoy observations are collected from the United States funded National Oceanic and Atmospheric Administration (NOAA) public data center. Although there exist many types of sensor payloads, we limit the scope to buoys with the Self Contained Ocean Observing Payload (SCOOP) [38]. Observations recorded on SCOOP buoys are transmitted via satellite to NOAA data servers for immediate access. Exactly 88 buoys are chosen for data extraction from a wide area of fixed locations that encompasses coastlines around the United States. Buoy sensors may be damaged, taken down for maintenance, or experience noisy local conditions. Each buoy measures multiple features per location, but uses individual sensors for each, leading to situations where only one feature may be missing. Missing values from the observations are interpolated, adding noise to the potential training data.

The numerically modeled ocean and climate models used are the HYCOM and ERA5 models, respectively [17][18]. The HYCOM and ERA5 data used are selected by finding the closest geographic and temporal resolutions. Both numerical models are grid-aligned, unlike the fixed locations of buoys, so the spatiotemporal alignment is not perfect. HYCOM is a higher resolution than the ERA5 data and typically fits the spatial position more closely as a result. Imperfect spatiotemporal alignment also introduces noise into the training pipeline. All data is combined to create a set of coupled and non-coupled training features. The complete set of features can be found in Table 1, for reference.

Table 1. Selected ocean features for training surrogate models. In bold are the numerical model features coupled with observations.

Feature Name	Feature Units	Feature Source
SST	°C	Buoy
Gust Strength	m/s	Buoy
Air Pressure	hPa	Buoy
SST	°C	**HYCOM**
Salinity	psu	HYCOM
Surf Elevation	m	HYCOM
Water Eastern Flow (U)	m/s	HYCOM
Water Northern Flow (V)	m/s	HYCOM
Wind Eastern Flow (U)	m/s	ERA5
Wind Northern Flow (V)	m/s	ERA5
Evaporation	m of w.e.	ERA5
Gust Strength	m/s	**ERA5**
Mean evaporation Rate	kg/($m^{-2}s^{-1}$)	ERA5
Mean Runoff Rate	kg/($m^{-2}s^{-1}$)	ERA5
Sea-Ice Cover (%)	[0-1]	ERA5
Air Pressure	hPa	**ERA5**
Cloud Cover	[0-1]	ERA5
Precipitation	m	ERA5

The data collected are from January 1, 2011, to December 31, 2011, and are taken in three-hour increments. The data is arranged into training, validation, and testing datasets by date. Training data are chosen from January 1 to September 13, the validation data is from September 13 to October 20, and the testing data includes the remainder of the year. Although one year of data is temporally small selection, this is exactly the same as what was used in the compared work [16]. The data is normalized based on the mean and standard deviation seen in the training data alone. The inverse is transformation upon model inference to investigate the results. Feature forecasts analyzed in the results section are in their respective scales.

B. Multiple-λ Physics-Regularized Loss

The physics-regularized loss function measures the surrogate inference error generated when comparing an observation value and the corresponding numerically modeled value. By evaluating the model inference against both sources of data, two error scores are produced for each observation and model pair. A ratio of the two error scores is taken, and this

error score is used for back propagation. The method is first proposed in [16]. The result is that, based on the ratio, the model is trained to approximate either source more strongly. The combination is determined by the hyperparameter λ, which selects a ratio of the errors to use. For example, in a forecasting task with features derived from sensor observations and numerical models, each feature's error is weighted by the singular λ value before being summed. This is proposed to improve the model by reducing the impact of interpolated or distorted values in either source.

Using a single λ value for all features is not optimal, as certain data sources could be more informative for specific features. To address this limitation, a more flexible loss function is explored in this work. Each feature is assigned its own independent λ value represented by a vector. This extension allows the model to assign different weights to each feature depending on the specific data sources being considered. The weighted errors accumulated from all features are then combined to calculate the total loss. Thus, features which display wildly different best-λ values are no longer required to use the same value. The subsequent piece-wise cost function can be calculated as follows in (1)-(6).

$$\Delta_{k,1} = |\hat{y}_{obs}^{(k)} - y_{obs}^{(k)}| \tag{1}$$

$$\Delta_{k,2} = |\hat{y}_{obs}^{(k)} - y_{model}^{(k)}| \tag{2}$$

$$\Omega_{\text{coupled loss}} = \sum_{k=1}^{K}[(\lambda_k * \Delta_{k,1}) + ((1 - \lambda_k) * \Delta_{k,2})]. \tag{3}$$

In (1) and (2), $\hat{y}$ represents the output of the surrogate model, while y represents the training ground truth and k represents an individual coupled feature. The source of modeling truth is determined by the subscript as y_{obs} or y_{model}. The error for each feature is weighted by λ_k before summing for the total coupled loss value. In this implementation, $K = 3$, which implies three coupled feature are included. Additional non-coupled features may be included in the surrogate and are defined as,

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

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

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. Additional numerical model features were added in this formulation, but no non-coupled observation features are included in the selection. Therefore, $\Omega_{\text{observation loss}} = 0$ in this implementation. Each piecewise value is summed into the final loss function (6),

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

This formulation of the loss function is similarly explored in an ongoing work [37]. However, the multiple-λ aspect is not explicitly explored, and the full impact is not seen. Also, the problem domain is significantly different, and does not compare to the original work. This paper expands on both papers by directly highlighting the multiple-λ physics-regularized loss approach and by comparing these results directly to the most related work.

C. Deep Learning Architecture

The prior work, [16], examines three neural network formulations, but the scope is limited in this paper to just one. Out of the three proposed architectures, the GRU model is chosen for further examination. Like the LSTM unit, a GRU-based model has weights which learn to store important context from the prior timestep. However, a GRU has one less internal gating signal and fewer weights as a result [33]. This yields a smaller model that is faster to train, which is important in this experimental setup, because many combinations of the λ vector must be iterated upon. Also, the GRU architecture was the one which benefited the most after using the coupled loss function in the previous paper, opening the possibility of continued improvement.

TABLE 2. GRU MODEL ARCHITECTURE. THERE ARE 24 TOTAL LAYERS WITH 1,827,306 TRAINABLE PARAMETERS. N REPRESENTS A VARIABLE BATCH SIZE.

Layer Type	Output Shape	Param #	Activation
Input Layer	(N, 18, 1)	0	None
Reshape	(N, 1, 18)	0	None
Dense	(N, 1, 256)	4,864	Tanh
Batch Normalization	(N, 1, 256)	1,024	None
Dropout	(N, 1, 256)	0	None
GRU	(N, 1, 256)	394,752	Tanh
Dropout	(N, 1, 256)	0	None
GRU	(N, 1, 256)	394,752	Tanh
Dense	(N, 1, 256)	65,792	Tanh
Batch Normalization	(N, 1, 256)	1,024	None
Dropout	(N, 1, 256)	0	None
GRU	(N, 1, 256)	394,752	Tanh
Dropout	(N, 1, 256)	0	None
GRU	(N, 256)	394,752	Tanh
Dropout	(N, 256)	0	None
Dense	(N, 200)	51,400	Tanh
Dropout	(N, 200)	0	None
Dense	(N, 200)	40,200	Tanh
Dropout	(N, 200)	0	None
Dense	(N, 200)	40,200	Tanh
Dropout	(N, 200)	0	None
Dense	(N, 200)	40,200	Tanh
Dropout	(N, 200)	0	None
Dense	(N, 18)	3,618	Tanh

The exact architecture of the surrogate model is found in Table 2. Dropout and batch normalization layers are added to prevent exploding or vanishing gradients during the training procedure. The Hyperbolic Tangent (Tanh) activation function is used for each layer. The model is trained for 100 epochs with a batch size of 256. The input and output vectors are the same shape to allow for a recurrent forecast style, where the forecast for time $t + 1$ uses the prior forecast from time t. So, only the first forecast is based on initial conditions. The model inference vector corresponds directly to the Table 1.

However, the only features considered in the ultimate analysis are those which are collected through buoy sensors, i.e., SST, gust strength, and air pressure. The GRU model conducts eight consecutive forecasts of three-hour steps to produce a daily 24-hour forecast for analysis. The model is trained using many buoy locations and is meant to be used as a generalized forecasting model for any buoy, although it only forecasts them individually.

D. Experimental Setup

To compare the proposed methodology with the existing approach, the same GRU network architecture, training procedure, and testing methodology as described in the literature is used [16]. The findings of the experimental setup are compared directly to the prior results. The best λ value was previously found through an extensive grid search to find the best singular hyperparameter value. This was described as a time-consuming process, and the time complexity is worsened by introducing three λ values instead. So, a linear grid search over the entire search space is impractical for this work. As a method to quickly validate the research question, a basic random search scheme is implemented.

The random search scheme is implemented by randomly generating λ combinations to control the coupling ratios in the physics-regularized loss function. This was completed in two phases. In the first phase, completely random values were used, and features were randomly chosen between 0.0 and 1.0 with a step size of 0.001. To this end 24 random permutations are evaluated. Secondly, the search space is narrowed for each feature such that SST is bounded between [0.500,0.990] and both gust strength and air pressure are bounded between [0.800, 0.999]. These values correspond to the regions where the previous paper saw performant results. A further 85 trials are run in this way. Each of the 109 trials are executed with the same random seed. Since the trials are randomly chosen, this is not guaranteed to explore a full range of possible λ combinations. However, the random ranges identified are sufficient for exploring whether a multiple-λ setup can produce more performant results, especially in the case of the secondary bounded search. To compare directly against the original results, the multiple-λ results are compared against the single-valued results between [0.0,1.0] with a step size of 0.1. The best single-λ values recorded for each feature are also compared. Comparing the multiple-λ and single-λ experimental results yields a total of 122 total comparative test cases.

It is notable that this technique does not scale well to problems without prior knowledge of the system, such as the one studied in this case. In the future, an efficient mechanism for discovering the best λ should be explored and ongoing research has for this task has already started [37]. However, justification that the proposed loss function is an improvement should still be given. Therefore, the aim is to show that some set of λ values exists which performs better than the previously found single λ value.

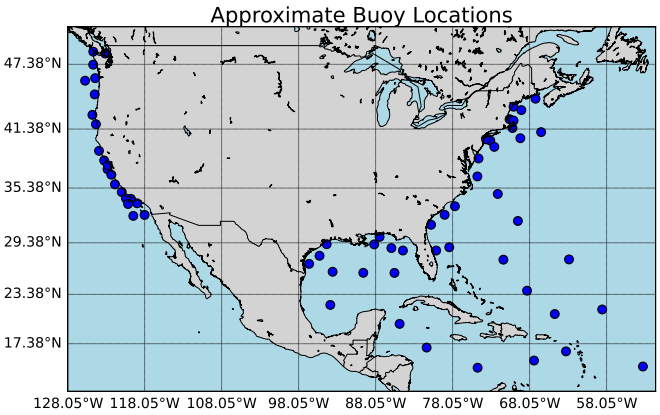

Figure 1. Approximate locations of the 88 buoys used in the testing dataset.

For evaluation of the proposed methodology, a testing dataset is composed of 48,039 instances taken from 88 independent buoys with an 8-step rolling horizon window. This window represents 24-hour forecasts. Each of these forecast windows is then evaluated and aggregated together. Evaluating the models on many buoys means the best surrogate is the one which is most accurate for a wide range of conditions. The approximate locations of each buoy is given in Figure 1, showing the diverse testing conditions. Given the forecasts, the Root Mean Square Error (RMSE) is taken for each coupled feature separately. The RMSE is defined in (7) as,

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

The parameter N is the number of test samples, y is the ground truth, and $\hat{y}$ is the prediction vector. Once the RMSE has been calculated for each feature, scores are summed together to determine which combination of λ values produces the smallest value.

IV. Results

An analysis of the most impactful results is illustrated in Figure 2 and further detailed in Table 3. Error scores are accumulated through eight forecast steps over a 24-hour horizon. Instead of displaying all experiments, only the top 25 results of the 122 λ combinations are highlighted. Among the top 25, only two are from single-λ experiments. The top two of those are ranked sixth and 23rd, respectively. Some single-λ experiments display minimal RMSE values for a single feature to the detriment of others, yielding a high summed RMSE. For example, the 45th ranked result is the best ever recorded RMSE for gust strength, while the SST RMSE is comparatively very poor. This shows that some selections of λ can minimize the test error of a single feature at the detriment of others. Although multiple λ values can still exhibit this behavior, increasing the hyperparameter search space allows more flexibility to choose λ values which

TABLE 3. TOP 10 PERFORMING COMBINATION OF λ VALUES ALONGSIDE EVERY SINGLE-λ BENCHMARK (BOLDED). THE RESULTS ARE SORTED BY THE SUM OF THE RMSE AND THEIR TOTAL RANK OUT OF THE 122 TEST CASES IS DISPLAYED.

Rank	SST λ	Gust λ	Air Pressure λ	SST RMSE	Gust RMSE	Air Pressure RMSE	Sum of RMSE
1	0.569	0.992	0.995	1.844	3.944	5.088	10.877
2	0.573	0.997	0.990	1.604	4.236	5.103	10.943
3	0.894	0.820	0.957	1.826	4.165	4.985	10.976
4	0.837	0.966	0.942	1.925	3.963	5.109	10.997
5	0.518	0.971	0.960	1.698	4.107	5.200	11.005
6	**0.900**	**0.900**	**0.900**	1.607	4.081	5.349	11.037
7	0.870	0.959	0.948	1.635	4.140	5.306	11.080
8	0.670	0.944	0.940	1.847	4.074	5.167	11.087
9	0.848	0.909	0.995	1.704	4.145	5.249	11.098
10	0.900	0.863	0.922	1.748	4.203	5.156	11.108
23	**0.960**	**0.960**	**0.960**	2.126	4.017	5.154	11.296
45	**0.840**	**0.840**	**0.840**	2.262	3.894	5.388	11.544
57	**0.800**	**0.800**	**0.800**	1.801	4.388	5.420	11.609
81	**0.700**	**0.700**	**0.700**	2.238	3.947	5.754	11.938
92	**1.000**	**1.000**	**1.000**	1.970	4.055	6.051	12.076
102	**0.600**	**0.600**	**0.600**	1.785	4.182	6.401	12.368
107	**0.500**	**0.500**	**0.500**	1.757	4.387	7.402	13.545
118	**0.100**	**0.100**	**0.100**	1.907	4.501	8.175	14.583
119	**0.300**	**0.300**	**0.300**	2.029	4.176	8.419	14.624
120	**0.000**	**0.000**	**0.000**	2.138	4.738	8.202	15.079
121	**0.400**	**0.400**	**0.400**	2.045	4.348	8.713	15.106
122	**0.200**	**0.200**	**0.200**	1.850	4.560	9.081	15.492

reduce error on average. This is shown in the fifth ranked result where individual feature do not perform better than in a corresponding single-λ setup, but overall improvements are seen. This is the main benefit of using multiple λ values instead of the previous methodology.

The top five results are all found from using a multiple-λ setup. This surpasses all previous outcomes found in the original findings. This suggests that using multiple λ values can enhance the hyperparameter space for improved test outcomes. The magnitude of improvement between the best single- and multi-λ setups is minimal overall, as displayed in Figure 2. The difference in error is more significant when compared to less optimal single-λ results. Most importantly, a set of independently selected λ values that yields better performance was found, satisfying the main goal. The consistency in the best performing λ configurations indicates that prior domain knowledge of the best-performing λ values is advantageous when conducting a random search. For example, although SST gives minimal RMSE results for a wide range of λ values, SST and Gust Strength both prefer a smaller range. Differences in numerical models influence the best selection of λ, for example, the spatial resolution is different in the HYCOM and ERA5 models. Specifically, ERA5 is a lower resolution than HYCOM, so individual grid points may be further away from the actual buoy locations.

In Table 3, the top ten results are compared with all single-λ results. Each row displays the rank out of all tests, the selection of λ for each feature, and the RMSE scores of each feature. One notable observation is that most single-λ setups are ranked worse than the top 25 results. The values for λ tend to be somewhat similar in the best performing results, depending on the feature. This behavior is in part due to constraining the random selections within previously found performant regions of λ for each feature. The λ choice for air pressure and gust strength benefit the most from prior knowledge of the optimal λ region. Specifically, for air pressure, the RMSE tends to be higher when $\lambda < 0.9$. Interestingly, the SST forecast achieves high performing results for a wider range of values. This lends credibility to the use of a random search setup when there is prior understanding of what λ values might be most effective. Also, this implies that the coupled numerical model highly influences the selection of best-λ value.

Although the sum of the RMSE is reduced when analyzing a multiple-λ experiment, individual feature results should still be considered. The top three single-λ results, are those which previously produced minimal error scores for one feature. Certain multiple-λ combinations yield lower feature-specific RMSE than those prior best results. For example, the third ranked λ configuration yields the lowest RMSE score for air pressure ever recorded using the demonstrated methodology. The second ranked result showed the best performance for SST ever recorded. However, a lower individual gust strength RMSE was never found, compared to the best performing single-λ result. The lowest sum of RMSE did not yield any best-result individual forecasts but had consistently low RMSE across all features. It is notable that a set of λ values which finds most performant forecasts for all features simultaneously was not found. This means that no single feature was optimized to the detriment of the other features. This suggests that using multiple λ values that are specific to each numerical model can overcome bias. This is because a single λ value is not allowing a biased numerical model to be more influential in the training process. Specifically, ERA5 has a lower resolution than the HYCOM data, which tends to mean that the HYCOM data is well fitted to the observations across all values. This describes why a broad selection of λ values work well for HYCOM, but not the ERA5 data. By using separate λ values for each feature neither numerical model source is forced to provide a suboptimal combination of data.

In Figure 3, absolute forecast error is highlighted. The error ranges over consecutive 24-hour cycles and is calculated based on the Mean Absolute Error (MAE) between the buoy derived observation and the predicted value. The numerical model (HYCOM/ERA5) is given as a baseline to compare against. Error is calculated from forecasts of a single buoy with the identification code 41009. The segment of forecasts analyzed are taken from period 40 to 120, demonstrating the error found in 11 forecast cycles. Compared are the best multiple-λ model and the best single-λ model, outlined in Table 3. Compared against both is the numeric error generated when comparing the buoy error to the numerical models HYCOM and ERA5. It is observed that the best performing model does not outperform the single-λ setup in all cases. In some situations, the stability of either model might be superior. The multiple-λ model is more stable on average and tends to experience less extreme fluctuations in the forecast. Occasionally, either surrogate model can outperform the numerical model, but

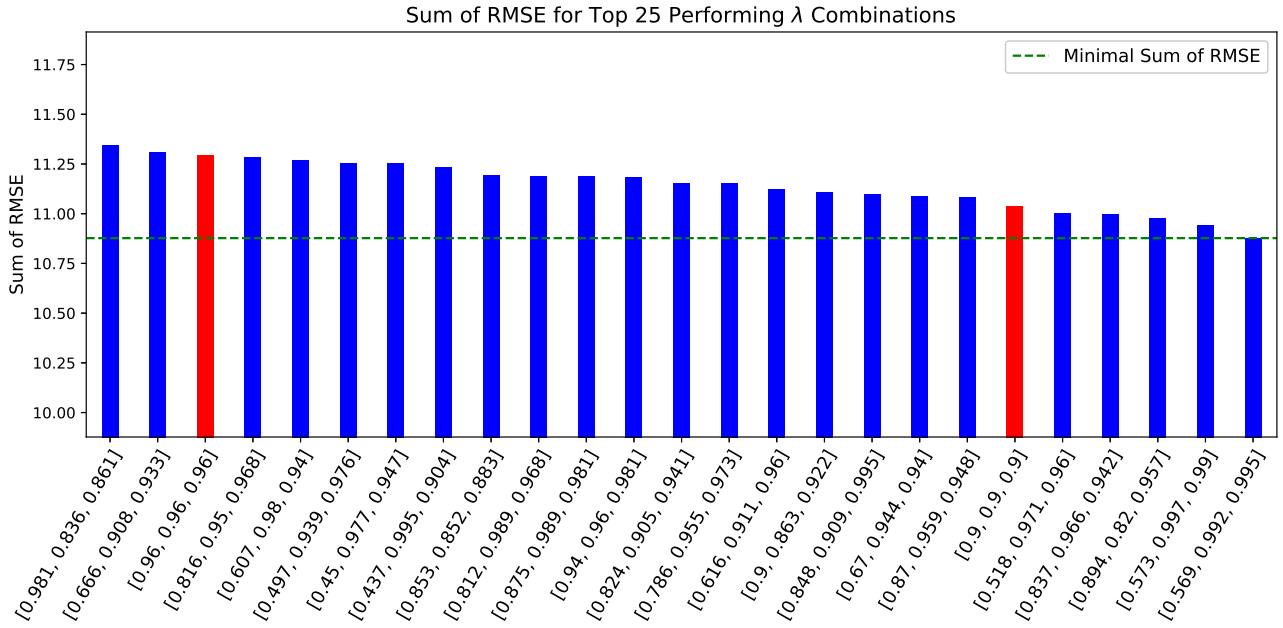

Figure 2. Top 25 performing λ combinations and their summed feature RMSE test scores. The green dotted line shows the separation of values between the most performant result and all others. Multiple-λ combinations are in blue, while any single-λ results are in red.

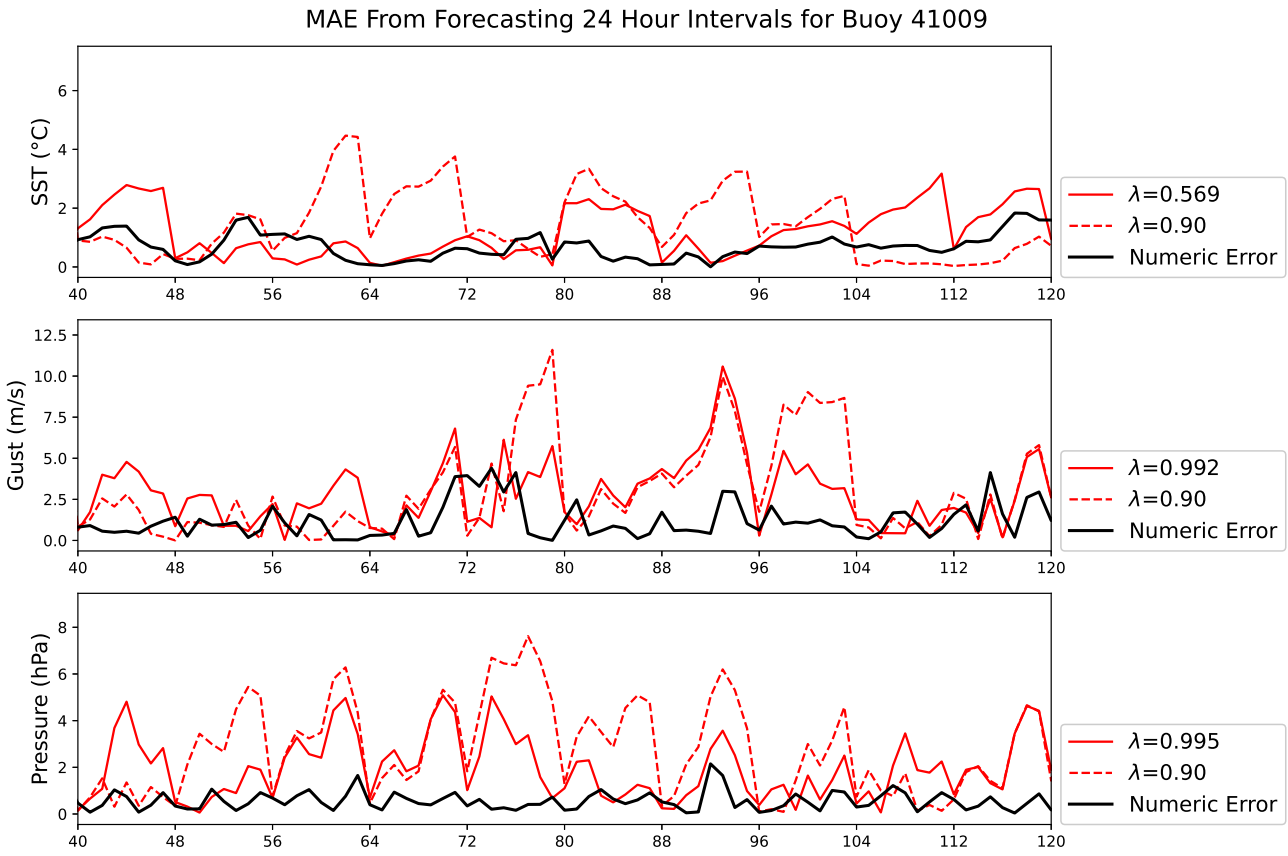

Figure 3. Shown are the forecast errors of the most performant combination of λ values compared against the most performant single-λ for a single buoy.

on average HYCOM and ERA5 show reduced error. This is expected, because statistical models are well-known to be less accurate than numerical models.

The trade-off in favor of the numerical models is the speed of the forecasts generated. Overall, the findings suggest that an independent selection of λ does improve the methodology by consistently reducing error across all features. Applying the correct weight for each ratio was shown to provide more a stable forecast on average. Finding continued improvements compared to the original research gives justification to the proposed methodology. The best λ values are highly dependent on the secondary data source (i.e., the numerical model) and, to a lesser degree, the selection of domain feature. Although a set of λ values which minimized the RMSE for all features was not found, that does not mean that a configuration does not exist. Exploration of the parameter space is the main limitation of these experimental results. More specialized search techniques should be implemented to efficiently find the best selection of hyperparameters. One further limitation of the methodology is the need for two sets of good-quality data. The benefits of the physics-regularized loss are directly dependent on the ability of the second source of data to be informative when the primary data, i.e., the observations, faces physical constraints.

V. Conclusion and Future Work

A previous methodology improved the forecasting of fixed-location ocean buoy observations by combining observation data with numerically modeled data. In the work, it was found that the selection of the ratio-determining hyperparameter, λ, varied depending on the numerical source and ocean feature. It was hypothesized that the results could be further improved if each feature was independently combined with numerical data. To address the proposed research question, the methodology was modified to include multiple independently selected λ values. The physics-regularized loss function was updated to combine features with numerical models in a less constrained way, which increased the potential hyperparameter search space. Then, a bounded random search was employed to generate random λ selections which produced superior results.

The updated technique was directly validated against the publicly available prior experiments. The outcome was a surrogate model that generated more accurate forecasts overall compared to the single-λ approach. The use of multiple λ values is particularly beneficial when multiple numerical models contribute to the feature set. For example, in this work both global HYCOM and ERA5 reanalysis models were used to improve overall results. A selection of λ values which reduced each individual feature's error below the best recorded value simultaneously was not found using the random search, but average error was improved for five combinations of λ. Such a combination of values may exist, even if the random search did not yield these results. It is acknowledged that the use of random or grid search to find the best parameter combination is time-consuming without prior domain knowledge and does not guarantee optimal results. However, the results justify the further use of multiple λ values, instead of a single value for all features.

Future work should validate the methodology using a wider range of real-world and theoretical datasets. Testing the combining technique with different combinations of input and output data would be very insightful. Different model architectures should be explored to assess the effectiveness of coupling data with more generalizable models. The physics-regularized loss is not reliant on the model architecture and should be attempted with more specialized architectures to see if similar improvements are found. Grid search and random search are not efficient enough, so developing methods for approximating or selecting λ values is a primary focus of future research.

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.

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] M. Angove *et al.*, "Ocean observations required to minimize uncertainty in global tsunami forecasts, warnings, and emergency response," *Frontiers in Marine Science*, vol. 6, p. 350, 2019.

[4] B.-T. Jong, M. Ting, R. Seager, N. Henderson, and D. E. Lee, "Role of equatorial pacific sst forecast error in the late winter california precipitation forecast for the 2015/16 el niño," *Journal of Climate*, vol. 31, no. 2, pp. 839–852, 2018.

[5] A. Bugała *et al.*, "Short-term forecast of generation of electric energy in photovoltaic systems," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 306–312, 2018.

[6] L. L. Lai *et al.*, "Intelligent weather forecast," in *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 04EX826)*, IEEE, vol. 7, 2004, pp. 4216–4221.

[7] H. Wang, Y.-M. Zhang, J.-X. Mao, and H.-P. Wan, "A probabilistic approach for short-term prediction of wind gust speed using ensemble learning," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 202, p. 104 198, 2020.

[8] W. A. Lahoz and P. Schneider, "Data assimilation: Making sense of earth observation," *Frontiers in Environmental Science*, vol. 2, p. 16, 2014.

[9] H. Y. Teh, A. W. Kempa-Liehr, and K. I.-K. Wang, "Sensor data quality: A systematic review," *Journal of Big Data*, vol. 7, no. 1, p. 11, 2020.

[10] W. L. Oberkampf, S. M. DeLand, B. M. Rutherford, K. V. Diegert, and K. F. Alvin, "Error and uncertainty in modeling and simulation," *Reliability Engineering & System Safety*, vol. 75, no. 3, pp. 333–357, 2002.

[11] Y. Tr'emolet, "Accounting for an imperfect model in 4d-var," *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 132, no. 621, pp. 2483–2504, 2006.

[12] G. Evensen *et al.*, *Data assimilation: the ensemble Kalman filter*. Springer, 2009, vol. 2.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] R. Bleck, "An oceanic general circulation model framed in hybrid isopycnic-cartesian coordinates," *Ocean modelling*, vol. 4, no. 1, pp. 55–88, 2002.

[18] H. Hersbach *et al.*, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[19] 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.

[20] 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.

[21] 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.

[22] E. Sandner *et al.*, "A multiple-location modeling scheme for physics-regularized networks: Recurrent forecasting of fixed-location buoy observations," *TechRxiv*, Aug. 2024. DOI: 10.36227/techrxiv.172469936.64312665/v1.

[23] E. Jang, Y. J. Kim, J. Im, Y.-G. Park, and T. Sung, "Global sea surface salinity via the synergistic use of smap satellite and hycom data based on machine learning," *Remote sensing of environment*, vol. 273, p. 112 980, 2022.

[24] G.-Q. Jiang, J. Xu, and J. Wei, "A deep learning algorithm of neural network for the parameterization of typhoon-ocean feedback in typhoon forecast models," *Geophysical Research Letters*, vol. 45, no. 8, pp. 3706–3716, 2018.

[25] 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.

[26] 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.

[27] 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.

[28] R. Niu *et al.*, "Multi-fidelity residual neural processes for scalable surrogate modeling," Proceedings of Machine Learning Research, vol. 235, R. Salakhutdinov *et al.*, Eds., pp. 38 381–38 394, 21–27 Jul 2024.

[29] 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.

[30] 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.

[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] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, IEEE, 2017, pp. 1597–1600.

[34] N. Thongniran, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Combining attentional cnn and gru networks for ocean current prediction based on hf radar observations," in *Proceedings of the 2019 8th international conference on computing and pattern recognition*, 2019, pp. 440–446.

[35] S. Wu, Z. Du, F. Zhang, Y. Zhou, and R. Liu, "Time-series forecasting of chlorophyll-a in coastal areas using lstm, gru and attention-based rnn models.," *Journal of Environmental Informatics*, vol. 41, no. 2, pp. 104–117, 2023.

[36] Y. Jiang *et al.*, "Prediction of sea temperature using temporal convolutional network and lstm-gru network," *Complex Engineering Systems*, vol. 1, no. 2, p. 1, 2021.

[37] 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.

[38] P. C. Kohler, L. LeBlanc, and J. Elliott, "Scoop-ndbc's new ocean observing system," in *OCEANS 2015-MTS/IEEE Washington*, IEEE, 2015, pp. 1–5.