5-2022

Forecasting Salinity in the Laguna Madre Using Deep Learning

Martin J. Flores Jr.
The University of Texas Rio Grande Valley

FORECASTING SALINITY IN THE LAGUNA MADRE USING DEEP LEARNING

A Thesis

by

MARTIN J. FLORES JR.

Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: Civil Engineering

The University of Texas Rio Grande Valley

May 2022

FORECASTING SALINITY IN THE LAGUNA MADRE USING DEEP LEARNING

A Thesis
by

MARTIN J. FLORES JR.

COMMITTEE MEMBERS

Dr. Jungseok Ho
Chair of Committee

Dr. Dongchul Kim
Co-Chair of Committee

Dr. Fatemeh Nazari
Committee Member

Dr. Jungwoo Lee
Committee Member

May 2022

ABSTRACT

Flores, Martin J., Jr., Forecasting Salinity in the Laguna Madre Using Deep Learning. Master of Science (MS), May, 2022, 48 pp., 6 tables, 12 figures, references, 43 titles.

Salinity is an important metric in the Laguna Madre for establishing the long term health of the local ecological population. By utilizing Deep Learning (DL) techniques, the predicted and forecasted salinity in the Laguna Madre is generated from data provided by the Moderate Resolution Imaging Spectroradiometer (MODIS)-Aqua satellite.

Currently, only one other DL model has been used to forecast Sea Surface Salinity (SSS), being a Recurrent Neural Network (RNN). However, the RNN model requires the prediction of a full area of salinity to function.

As such, several model architectures were tested, with the best one, being a Multi-input MPNN, utilized to evaluate the feasibility of forecasting utilizing simpler DL models. The results show that a one-day forecast is plausible, while three and five-day forecasts would require a data-rich environment, unlike that of the Laguna Madre.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Background

The Laguna Madre is one of few hypersaline estuaries in the world, providing a unique habitat to many different creatures. Prior to dredging the Gulf Intracoastal Waterway, salinity commonly reached values above 100 ppt, frequently causing the death of many local fish and wildlife [1].

Given the unique environment provided by the Laguna Madre, it is key to provide adequate monitoring solutions for the area. Particularly, given the history of the area in regard to salinity, it is key to monitor this metric in order to ensure stability of the environment for wildlife of the area.

However, setting up monitoring stations and actively sampling salinity is costly, and requires a great time commitment. As such, predicting Sea Surface Salinity (SSS) from remote sensing data has been of interest for a while.

It's only in the past decade that techniques, along with the technologies used in remote sensing have greatly improved [2]. The recent shift from numerical modeling to models utilizing machine learning improved the accuracy and resolution of results.

Advances in techniques of image processing helped recover more images from satellites, and new methods for extracting valuable data from processed images more accurately captured what researchers were trying to gather from the data.

In general, SSS is an important metric in modeling biological environments [3]. With the Laguna Madre having a unique biological environment, fostering species specific to the area such as unique microbial communities [4], the importance of modeling and forecasting salinity is key to the stability of the unique environment.

Current research on the application of machine learning in remote sensing data, particularly satellite based data, tends towards the recognition of various attributes from the data. Whether it be pCO2 [5, 6] or SSS [7, 8], the primary goal of these studies is to create a model, using machine learning or Deep Learning (DL) techniques, that has the ability to estimate the selected attribute from satellite imagery. However, there has been little attention given to producing models that can create forecasts from satellite imagery.

Currently, there is a very limited amount of research done into the forecasting of Sea Surface Salinity utilizing Deep Learning techniques, particularly in the Texas area. One paper has created a model that achieved an accuracy between 99.29% and 96.85% providing one, five, and fourteen-day forecasts [9]. A similar technique was utilized to forecast Sea Surface Height Anomaly, though to a lesser degree of success with a seventy-two-hour forecast showing an average accuracy of 79.99%, with further forecasts being increasingly inaccurate [10].

By developing multiple models, using both a Convolutional Neural Network (CNN) and Multi-layer Perceptron Neural Network (MPNN) architecture, the better model for predicting SSS in the Laguna Madre can be discovered. By using the best model as a base, it can then be trained to then create a forecast for the Laguna Madre.

Thereafter, the models would be feed preprocessed data retrieved from satellite imagery into the model. To create forecasts, the ground truth points to train the model would have to be selected at the appropriate points in the future. A forecast for SSS in the Laguna Madre could then be created. The output from this model can then be used for planning studies or forecasting and understanding the effects of a storm event.

Complicating this study is absence buoy data on Sea Surface Salinity in the Laguna Madre. From buoy BZST2 at Brazos Santiago Pass to ANPT2 at Aransas Pass, there is no SSS data provided. Sea Surface Salinity data from ship surveys are limited in utility, given they only cover a small area, and don't have a large range of temporal data. This in turn limits research on SSS to the areas that are chosen for surveys. This however, does not give a complete picture of the area, nor will it tell how the salinity may be in the coming days.

The benefits that are provided by forecasting with satellite imagery data is great. While modern satellites have radio sensors that measure SSS directly, the resolution of the data ends up being approximately 60 km for reduced noise data or within the range of 40 km for noisy data. This, compared to image derived SSS, is significantly less that what could be produced using a Deep Learning based model. A DL based model, with the appropriate data, can reach a resolution of approximately 1 km. This allows for detailed observation of specific areas compared to the more global view that radio measured data provides. Imagery data, as apposed to point sampling data, provides a more complete view of the local area than simple models on select points.

Purpose and Objectives

The purpose of this research is to create a Deep Learning model to produce daily forecasts of Sea Surface Salinity within the region of the Laguna Madre. Along with the creation of the model, a dataset focusing on the Laguna Madre will be created, allowing for the development of the models highlighted in this paper. By creating a model of the area, the local environment could be better understood and allow for improved modeling of the Laguna Madre area.

As a sort of baseline, an existing numerical model, being TxBLEND, will be used for comparison. By utilizing an established model, the strengths and weaknesses of each method can be more thoroughly evaluated. Given that the Machine Learning based model will have different inputs and outputs, along with internal differences, there is some merit to briefly describing the functions of each.

TxBLEND, having been calibrated and validated in the past [11], extends just past the Corpus Christi bay. TxBLEND requires several parameters such as BigG for the wave continuity equation, Manning's n roughness coefficient to represent friction on the bottom, boundary conditions for the model's edge, tide data at the boundry conditions to aid in simulation, river inflow data, and meteorological data to be set in order for the simulations, such as wind stress, to be run.

The time series of the input data can vary, with some data needing to be hourly and others, a daily average. The output will be a triangulated mesh, with each vertex having its own value. One

clear advantage that TxBLEND has over DL based forecasting models is that, given enough data is gathered, the simulation can be run to predict salinity change on an hourly basis. The forecasting model that will be developed is largely limited by the time it takes for a satellite to pass, which at best has a one-day forecast.

The Deep Learning based model, compared to the TxBLEND model, will only require a gridded matrix of satellite data as input, with possible addtional gauge data, and will provide a gridded matrix of salinity data as the output. All other parameters are build into the model by selecting the model's structure as well as several hyperparameters such as batch size, convolution filter size, number of convolution filters, and learning rate to list a few.

The DL based model will also require less user input and be overall less complex. Once completed, the model will only need some Level-2 data from satellite images, and optional gauge data if built with it, for forecasting. This can be further simplified by automating the retrieval of Level-2 satellite data with a script. This is much more simplified compared to TxBLEND where several parameters need to be set for the model to be run, with a great deal of information being required to simulate the area.

Methodology

There are several steps that need to be made in order to create a forecasting model for Sea Surface Salinity. Before the model can be constructed, the input dataset must be created. The creation of this set is not a simple task, and requires some preliminary research to identify key attributes for estimating SSS to save computational time. Then, after the input dataset is constructed, the first step to creating the forecast model is to decide on the architecture of the model. Thereafter, is to optimize the model. Finally, the model must be tested for its accuracy. By following these steps, the DL based forecasting model can be created.

All the models that will be developed in this paper require validation. Not only do the models require validation, but the data itself requires some degree of quality checking. In order to verify the output of the model, several gauge stations and cruise surveys will be used as ground truths to

then be compared to the output of developed models. Should the accuracy of the model's output be below the expected threshold, alternative sources of data, including additional inputs, should be considered.

When deciding how to best create the dataset, whether it be the training or test dataset, several things must be considered. The purpose, attributes worth keeping, and the size of the dataset are of concern. In this case, the dataset will be focusing on the Laguna Madre area, with data on SSS for each day if possible to create as complete an image as possible.

As a starting point, Level-2 Ocean Color data along with Sea Surface Temperature (SST) and location data will be collected from times matching the available SSS gauge and cruise data within a reasonable margin, being ± 6 hours for matching with cruise data and an exact time match for gauge data. All possible data, being gauge and ship survey data, for the selected time period will be gathered for a higher chance of matching quality satellite data. By utilizing this combination of remote sensing data as well as gauge and ship survey data, high resolution SSS maps can be created to then be utilized for analysis.

After the initial gathering of data, some time and care must be put into the cleaning of the dataset, as only some portions of the dataset will likely be unusable for the models [8]. However, some lower quality data may still be valid for the purpose of making the model more robust to noise. The data will be retained for use with this in mind. If the lower quality data is deemed unusable for this purpose, then data will be removed from the set.

All the developed models will extract data from six key attributes provided by the satellites. The input for this simpler models will be five different Remote Sensing Reflectance (Rrs) bands provided by the satellites as well as the Sea Surface Temperature (SST) measurements from the satellite. The inputs are then normalized so the model will be able to treat the inputs in an unbiased manner [12].

The extended dataset with gauge data will be used to train the Multi-input variants of the model, which just like the previous models, will produce a high resolution SSS map. Once a reasonable accuracy has been obtained from the models, they will be tested on a portion of the dataset reserved

for testing. The model with the best accuracy will then be used to forecast SSS in one, three, and five-day intervals. These intervals were chosen to provide sufficient testing for various extended outlooks without extensive resource requirements.

The several models will be developed using various techniques, such as using dropout [13] to prevent over-fitting of models, and using a hyperparameter tuner to optimize what architecture and attributes would be the best. Several iterations of tuning will occur, allowing for the models to be further tuned and various configuration of the model to be tried. The tuned version of the model will then be run on the test set to then be compared to the other models.

The forecasting model will be created with the goal of performing a one, three, and five-day forecast of SSS in the Laguna Madre. The lengths of these forecasts allow us to estimate how viable intermediate and further forecast are.

The unit that will be used for salinity will be Practical Salinity Unit (PSU). The other significant mode of measurement commonly seen for SSS is the Parts Per Thousand (ppt) measurement of salinity. However, PSU is much more commonly used given the ease of measurement and the standardization of the processing of data. This comes at the cost of precision, in which, the ppt measurement would be much more accurate.

The resolution of the final model is dependent on the quality of the given satellite input image. In optimal conditions, resolutions of approximately 1 km should be achievable, whilst other conditions may lead to overall lower quality images. These conditions are an artifact of satellites, given their specific orbital patterns leading to differing angles of sensor readings.

CHAPTER II

PREVIOUS STUDIES

Deep Learning Primer

In order to gain a deeper understanding of how Deep Learning works, as well as how Deep Learning models can extract information from various data sources, a brief primer of Deep Learning is in order. Generally speaking, the idea of creating computational neurons, based off our understanding of neurons in neuroscience based studies, has been around since at least the 60's [14]. For the most part, it was being held back by the lack of computational power at the time.

However, in recent years, there has been a boom in the field of Deep Learning. Given the relative accessibility of the Graphics Processing Unit (GPU), research on Deep Learning became powered and accelerated by GPUs. Further developments in the acceleration of model training can be brought about with a Tensor Processing Unit (TPU), however, that is out of the scope of this paper.

MPNN

Going into the details of how Deep Learning models are powered, every model is fundamentally powered by what is called the artificial neuron. This, in turn, is modeled by a multiplication by the input's weight, which relatively corresponds to the neuron's importance, the addition of a bias term, and a summation of the inputs.

In order to generate non-linear activity, a mathematical function, termed an activation function, is placed at the end of the artificial neuron. This activation function was originally chosen to be the sigmoid function, as that most closely represents our own neuron's activation. However, it's

important to note that there are a slew of activation functions, and it is ill-advised to simply use sigmoid without testing other activation functions. An example of an alternative activation function is the ReLU function. This function is linear from exect in cases where the inputs in negative, in which the ouput is zero.

Given that these neurons simply work on any input and provides an output, the next natural step to further extracting the power of neurons would be to create groups of neurons. These groups in turn can also take inputs from the previous layer and then processes them on further. This led to the Artifical Neural Network (ANN), which is generally equivalent to the Multilayer Perceptron Neural Network (MPNN) and the Neural Network.

These layers themselves have their own name as well, with the first layer being the input layer, the last the output layer, and the rest called the hidden layers. Of interest is that as long as you have one hidden layer, if that hidden layer were to have an infinite number of neurons, it would be possible to approximate continuous functions [15]. Of course, this is just one interpretation of deep learning neural networks.

Once the foundations were laid and the machinery was available, Deep Learning became a popular field of research. Many focused on how different architectures, referring to the number of hidden layers and neurons per layer, affect the results. However, there were limitations within the plain neural network architecture itself. These limitations would lead to the branching of models architectures, allowing for neural based models to be applied in other areas.

CNN

Given the non-linearity of deep learning models, it has a great capability for the processing of data. However, there was some difficulty with creating models that could deal with complex images well. Even gridded data could only produce good results if the data was relatively simple. Convolutional Neural Network (CNN) models can be thought of a specific extension of MPNN models, in that most CNN models actually use fully connected layers, a type of neural layer, to then process the data that the CNN has produced.

A typical CNN model is structured as having a convolution layer or layers, followed by some pooling layers if the dimensionality is too high, with some fully connected layers at the very end to process the data. The convolution layer works by sliding a filter, who's attributes a typically randomly generated, over the input data and performing a specified operation on that data. Each filter then produces an output based on a multiplication of its values and the input, followed by a summation of the previously formed product.

The idea behind CNNs is to extract further information from data that is gridded or image-like, such as satellite data from MODIS-Aqua. By further extracting information like this, patterns from the image can be extracted and the amount of information from gridded data is further increased compared to a simple MPNN. A simple MPNN would not be able to extract such relative information, as it process each point of data individually, compared to the batch processing of CNNs.

Other Models

There are, of course, further deep learning models of interest. Of particular interest is the Recurrent Neural Network (RNN) model. The RNN model is built to have a portion of the network feed back to itself, typically in a delayed fashion. This fed back bit of information is what allows it to be used for the next prediction. RNN models are effective for time series data, however, require data in a specific format to be utilized to its full effectiveness.

That isn't to discount the existing traditional statistical methods, as in most cases, they can be more than enough for predicting current salinity conditions. As will be shown shortly, traditional statistical methods can provide more than adequate results.

Literature Review

While not a model, SMOS and Aquarius are satellite missions with a common purpose being the detection of salinity directly from orbit using L-band radiometry. However, they are generally shadowed by other salinity estimation techniques as their visitation frequency is low, and the resolution is much lower compared to other more recent methods [16, 17, 18, 19].

There have several techniques used for predicting SSS in the past, such as multiple regression, linear regression, and principal component analysis (PCA)[20, 21, 22]. However, most of these techniques have fallen out of favor for the relatively new deep learning techniques. The reason being that more complex areas, as well as large swaths require more complex models to handle the local variability of the area.

In terms of forecasting models, the Recurrent Neural Network (RNN) based solution developed by Song et al. [9] allowed for daily average salinity forecasts to an arbitrary date. The sort of RNN model used within the paper is quite powerful, allowing it's on output to be used as input, and capable of learning trends with respect to time [23].

Although in this particular case the RNN model was trained to produce daily forecasts, the model can be trained to produce arbitrary timescale forecasts, with the main limitation being that training data within the same timescale is necessary to produce accurate models. However, even with such a powerful model, it was noted that there was difficulty creating accurate forecasts when rapid salinity fluctuations were present.

Also of note is the data used within the paper. The dataset, being Reanalysis dataset of the South China Sea (REDOS), is a reanalysis dataset for the South China Sea. It has approximately a decade worth of daily data, at multiple depths.

What this means for the model is that it's built to take a salinity state as input and output a new state, which requires a map of salinity to be generated for the entire sea before a forecast is created.

Table 2.1. Various methods used to predict SSS, and respective results.

Author Year	Model Type	Data	Model Results	
			RMSE (PSU)	R^2
Song et al. 2020	RNN	REDOS	0.1294	-
Chen et al. 2017	MPNN	MODIS-Aqua, SeaWiFS	1.2	-
Schoenbaechler et al. 2011	Numerical Model	TWDB, TCOON, UTPA	0.7*	-
Marghany and Hashim 2011	Box-Jenkins	MODIS-Aqua	-	0.98
Urquhart et al. 2012	ANN	MODIS-Aqua	2.50*	-
	GAM		2.38*	-
Zhao, Temimi, and Ghedira 2017	MLR	OLI	1.87	0.6
Geiger et al. 2013	NN	MODIS-Aqua, SeaWiFS	1.85	-

* Denotes measurements using ppt.

This is precisely what allows the model to be run repeatedly on its own input to generate arbitrary length forecasts. As such, while the model is very powerful, it cannot be used as it is and further research would need to be done to fully utilize this model.

While the RNN model is extremely powerful, it does require a specific data requirements that not all areas are able to meet. In this case, a simpler model would be a better choice. The Multilayer Perceptron Neural Network (MPNN) model developed by Chen et al. [8] has shown that predicting salinity within the Gulf of Mexico is, indeed, feasible even with a sparser dataset. This is of particular interest as the study covers a large area of the northern Gulf of Mexico, which is the closest study geographically to what is being proposed.

This MPNN model was trained to predict salinity conditions at the time that the satellites fly overhead and record their data. This model, however, doesn't have the advantage that advanced DL models have, such as the RNN model discussed earlier. That is, this model can only provide a single output, unless trained in specific manner, and the output is such that the model cannot run again on it's previous output alone. Simply put, the model can only predict a single timestep based on its input.

Although there can only be a single timestep predicted, the limitations provide an advantage. The input dataset can be much more sparse in both spatial and temporal dimensions, compared to the previously mentioned RNN model. For areas with low data availability, such as the Laguna Madre area this type of model is particularly important. It allows for the prediction and forecasting of SSS without the major data requirements of more powerful models.

Of note in this study is that there was some difficulty in predicting the salinity within areas that had strong upwelling and algal blooms. The proposed reasoning for this is that the two previously mentioned events cause a depression in the blue Rrs bands, leading the model to predict that there is lower salinity within the area.

An interesting statistical method was used by Marghany and Hashim 2011 [24]. This statistical method, being the Box-Jenkins method, is a type of ARIMA model which allows for the forecasting of values based on three types of models, being autoregressive, moving average, and a combination

of the two previous modes. In this particular case the model was tuned to forecast monthly averages of salinity, allowing for the modeling and forecasting of seasonal changes within the area. This allows for improved accuracy by effectively removing significant variance from the dataset.

The data used within this study is from the MODIS-Aqua satellite, with the area utilizied being largely focused around the East coast of Malaysia. This particular model requires a deal of preprocessing to transform the data into appropriate input for the regression model.

As for numerical models, one of the few models set for the Laguna Madre is TxBLEND. Having been calibrated for the area [11], its performance is currently the best within the area and has been actively used by the Texas Water Development Board (TWDB) to simulate conditions within the Laguna Madre estuary, as well as other areas for various projects that the TWDB require.

TxBLEND works to model water circulation and salinity transport, and does so by extending it's predecessor, the BLEND model, and augmenting it taking advantage of numerous inputs. The inputs added to the model were river inflows, tides, salinity, winds, evaporation, and various other routines [25]. By adding these several additional inputs, the data requirements became much higher, however, the model's performance also increased.

Having been around for a number of years, TxBLEND has an established number of advantages over more modern modeling techniques. There's existing data for TxBLEND to be based off of, and if not, there are several established techniques to help fill missing data, both spatial and temporal, for numerical models. These techniques then allow for a wide range of data from various sources to be utilized, improving the performance of the model furthermore.

Numerical models such as TxBLEND also have the advantage of ease of forecasting, by easily using the new output of the model itself. While self-feeding DL based models similar to this are possible, the input data varies. As such, the data requirements differ, and are not exactly directly comparable.

A very in depth paper by Urquhart [26] shows the utility of statistical methods, while still highlighting the powers of ANNs. Within the paper the following models were utilized: Generalized Linear Model (GLM), Generalized Additive Model (GAM), Artifical Neural Network (ANN),

Multivariate Adaptive Regression Spline (MARS), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Bagged Categorical and Regression Trees (BCART), Random Forest (RF) trees, and a mean model for a statistical null model. This variety of models is selected to compare existing statistical models to DL based models.

The data utilized within the study is from the MODIS-Aqua satellite, with salinity sample data taken from the Chesapeake Bay Monitoring Program, along with cruise samples obtained from the Maryland Department of Natural Resources (MDDNR) and the Virginia Department of Environmental Quality.

The in-situ measurements were then selected based off of the mean optical depth, dropping values that were sampled at deeper depths. As for the satellite data, the data was processed using the SeaWiFS Data Analysis System (SeaDAS) and reprojected into a cylindrical coordinate system. Thereafter, quality flags were used to drop invalid data points.

Although the paper goes over eight models, with an additional null model, only the two best performing models based on RMSE over the entire validation set will be discussed for the sake of brevity.

Starting with the better performing model, the GAM could be considered an extension of the GLM, in which the GAM allows for a non-linear relationship between the independent variable and the dependent variables. The way it achieves this is by providing a smoothing function to allow for a non-linear relationship. In this particular case, the function used was a cubic regression spline.

As for the ANN, the model was based off an existing study in the area [27]. In this particular case, two models were trained using the same techniques, with the difference being the number of neurons within the hidden layer, being 40 and 45. From there, the model with the best training score was then selected to be tested on the validation set. Noted was a dependence on the randomly initialized weights, which is an inherit property to ANNs.

In the testing of models, a comparison of the base dataset, as well as the geographically augmented dataset were compared to see whether or not the additional information provided by the coordinates would be significant. Upon analyzing the results using the GAM, it was determined

that latitude and longitude were statistically significant to the model. In fact, out of all the parameters tested, the most significant values were the latitude and longitude values.

Taking a look at an interpretable model, the Multivariable Linear Regression (MLR) model is a model that focuses on utilizing Rrs to predict the concentrations of SSS.

In this particular case, the data was retrieved from two sources. Rrs data was retrieved from the Landsat-8 satellite, specifically the Operational Land Imager (OLI) with the SSS data being collected from 52 stations over several missions from June 2013 to November 2014.

One of the great advantages of simpler models such as MLR is the interpretability of the model. It allows for a further analysis of the relationships of the parameters and for causal analysis utilizing the model as a basis for the understanding of the relationships.

Many of the past models that utilize data from satellites tend to use both sea surface temperature, remote surface reflectance, as well as chlorophyll-a concentrations [20, 21, 24, 26, 27, 28, 22]. The reason for the primary use of these variables has to do with what correlations could possibly be found from space.

Interestingly, Colored Dissolved Organic Matter (CDOM) is a good indicator for salinity that remote surface reflectance can pick up on [29, 30, 31, 32]. Given the previous justification, CDOM absorption coefficient (a_{CDOM}) can read using ocean color measurements, and thus, be used to predict SSS concentrations using satellite data [33, 34, 35].

Of note is that the use of CDOM to predict SSS is based off the assumption that there is conservative mixing. Furthermore, the relationship between a_{CDOM} and SSS can vary based off of location, with some areas showing a linear correlation [36, 37, 21], and others showing much more complex relationships [30, 38, 39].

With this, the core of this paper, as well as several previous paper's base is layed. The previously mentioned processes allow for the Deep Learning (DL) based models to predict and forecast salinity. In more complex scenarios, the underlying assumption may be false. As such, model accuracy may decrease. The benefit of utilizing a DL based model in this case is to adapt to such cases and all for the retention of model accuracy.

CHAPTER III

MODEL IMPLEMENTATION

Overview of Data

Deep learning, and more broadly, machine learning, requires a great amount of data in order to generate accurate models. Since the recent boom in machine learning, there have been several techniques developed to help stifle these limitations, allowing for complex models to be developed with limited datasets. Still, there needs to be great consideration in what goes into a model. Biases and errors within the dataset itself translates to biases and errors within the model that the dataset is fed into.

The following section will cover two major datasets created for the model: Salinity for model validation and satellite data as input. Discussed will be sources that the data is gathered from along with the details on the data either discovered by analyzing the data or by reading accompanying documentation.

Salinity Databases

One of the first data sources searched for salinity data was Buoy data. Searching through buoy data provided by National Oceanic and Atmospheric Administration (NOAA) in area the Laguna Madre as well as nearby the coastline, it was found that none of the buoys, including the ones within the Laguana Madre, had any Sea Surface Salinity data.

Although the NOAA buoys didn't have any salinity data, the World Ocean Database (WOD) by National Centers for Environmental Information (NCEI) did. Salinity data has been gathered from the NCEI WOD dataset [40], with all but one cast of the 6142 casts being valid. After inspecting

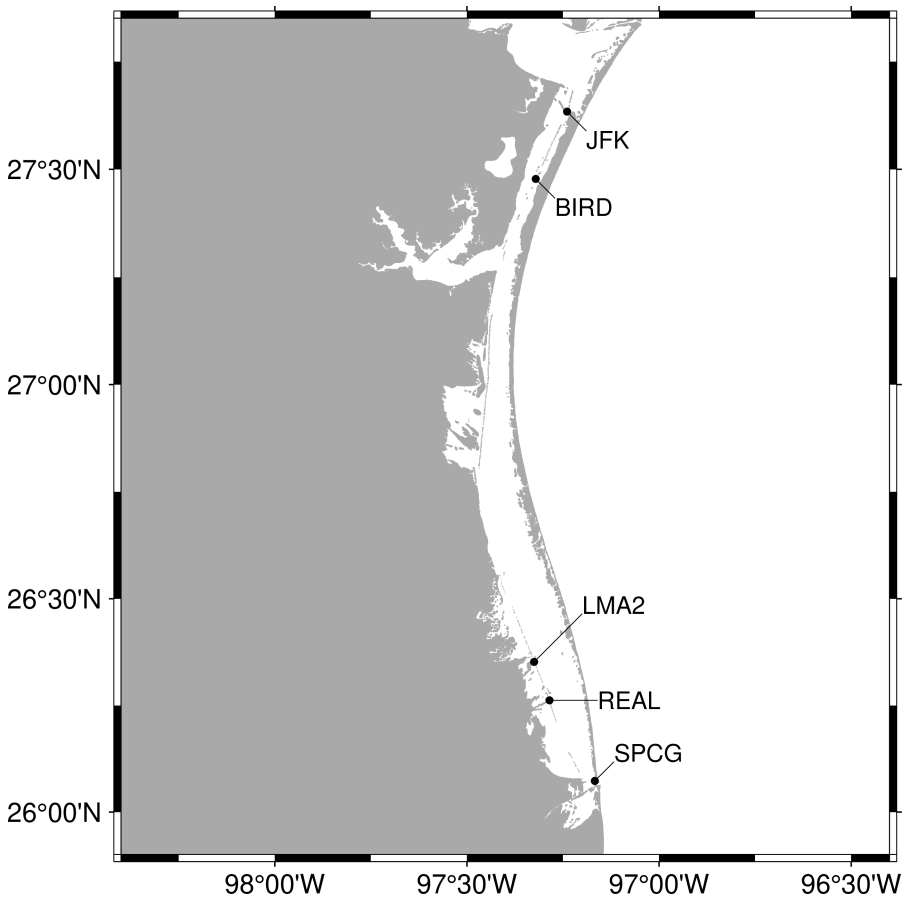

Figure 3.1. Long-term sample points of the Laguna Madre from the TWDB.

the data point carefully, a solution to fixing the one error allowed for the entire dataset to be used. The salinity data was gathered from the years of 2002 to 2020. The bounds for the search are a longitude of -98.93 to -93.05 with a latitude of 30.96 to 25.15.

Each cast can have several depths, however, each cast is not guaranteed to have the same depth as the previous. Every cast in the WOD dataset can have salinity measured at various depths, and not all the casts have each of the depths covered by others. This in turn reduces the number of available casts. Choosing a fairly common depth, such as 5 m, returns less than the total casts. However, when retrieving the data, there is the option to normalize the data to standard depths. As such, this processing option was utilized when retrieving the dataset.

Additional salinity data was retrieved from the Texas Water Development Board (TWDB)'s Water Data for Texas (WDFT). This resource provides long term sample points for several loca-

tions throughout the Laguna Madre, with multiple attributes to choose from. In this particular case, only salinity data was gathered. Mulitple gauge stations were selected to contribute to the training, and calibration of the Deep Learning model. The stations selected are, from northernmost to southernmost: JFK, BIRD, LMA2, REAL, and SPCG. These points, shown in Figure 3.1, are some of the few that had salinity data available for use.

While not strictly salinity data, several other data points were gathered from NOAA's gauge stations. Particularly, gauges nearby the TWDB gauge stations were selected such that the distance is minimized and the data is likely to have higher correlation to the measured salinity. The data points selected were wind direction, gust, speed, and barometric pressure. This attributes were used in model training and testing to see if they improved model performance.

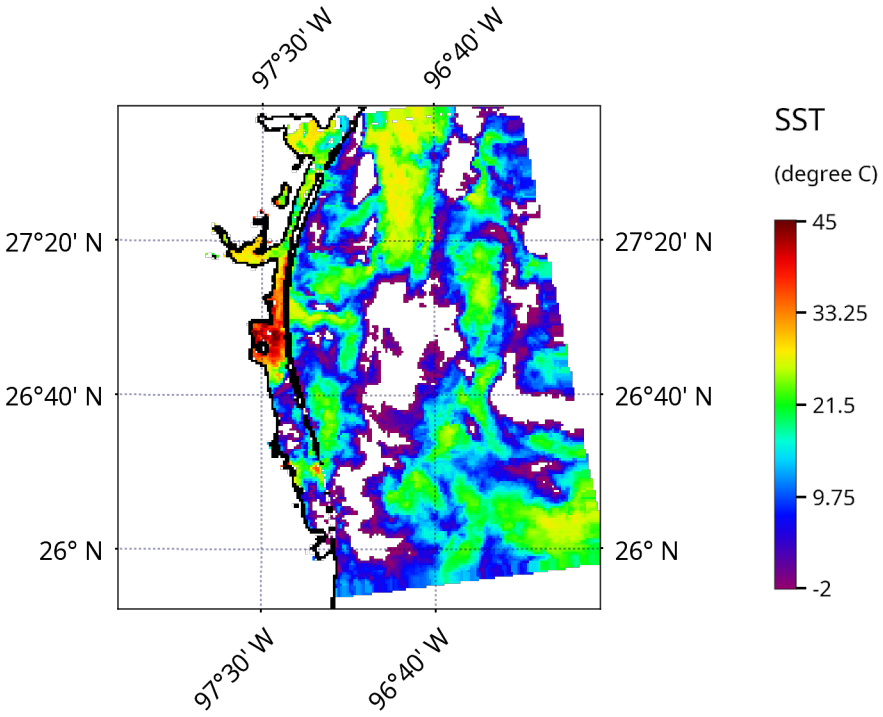

Figure 3.2. MODIS-Aqua SST Image from 2002-07-05

Satellite Databases

The choice to use the MODIS-Aqua satellite was made because it had the most data from mission start to end, with the satellite itself being still active. Additionally, the MODIS-Aqua satellite

has a relatively high resolution, allowing for fine gridded output. This allows the model to be used on the same mission and data it was trained on, as such, errors produced from the model results should be consistent unless the preprocessing in the input data is changed in the future. Both the Inherit Optical Properties (IOP) [41] data and the Ocean Color (OC) [42] data were used as model input.

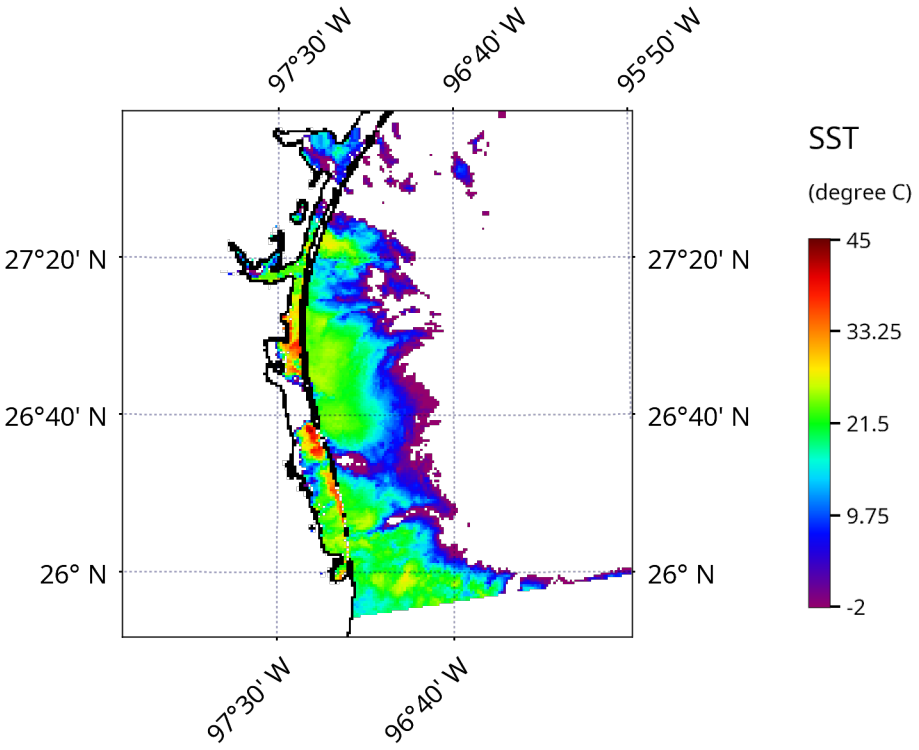

Figure 3.3. MODIS-Aqua SST image from 2002-08-08

The criteria for the satellite data are rather strict. The data must completely cover the bounds discussed earlier in the salinity data. This choice was made for the simplicity of point extraction, as attempting to discover whether or not the satellite data covers the salinity data can be avoided by simply ensuring the entire area of interest is covered.

The time period of the satellite data that was selected was set to be within the same month of an existing salinity data point. This means that the current satellite data is within the 2002 – 2020 range, with some months being dropped due to no salinity data existing in that point in time, predominately being the months of January, February, November, and December. An additional limiting factor is the additional gauge data if utilized.

While not pursued in this study, other satellites can be considered for additional input. With the current selection criteria for the 2002 – 2020 time period, there are plentiful data points to begin working with. However, additional inputs from other satellites could serve to augment data, by providing different samplings, regardless if it is sampling the same salinity point. One thing to consider whether augmenting the dataset with more images from another satellite would produce any benefit, particularly to input noise.

Additional satellite inputs could potentially allow for more of the salinity dataset to be utilized. However, if this is not done with some care, this could lead to duplicate entries that could impact the model's performance. Plus, there could be more biases incurred from the ratio of satellite images feed, as well as how each satellite data is processed and potentially transformed in some cases.

Another detail worth noting is that not all the Rrs bands that MODIS-Aqua provides are being used. It may be worthwhile trying the extra bands to see if there is any more performance to be gained at the cost of longer computation and processing times.

The data from the MODIS-Aqua satellite is of a high enough resolution, being approximately 1 km, that the images show the Laguna Madre. However, not all the images are going to be of high quality. Figure 3.2 and 3.3 show what some satellite images of lesser quality look like. While the clouds in Figure 3.2 and 3.3 may affect the forecasting ability for the Laguna Madre, there are still some areas than can be used for forecasting.

Data Analysis

To better understand the data that is used within the model, an analysis of the input data needs to be performed. This is to both apply data cleaning and transformation techniques to get the data into working order. Failing to perform these tasks would result in a model that replicates poorly optimized data. The analysis will be broken into two parts, for analysis of both the salinity data and satellite data. This allows for separate checking and transformation of the data, ensuring the quality of the data beforehand.

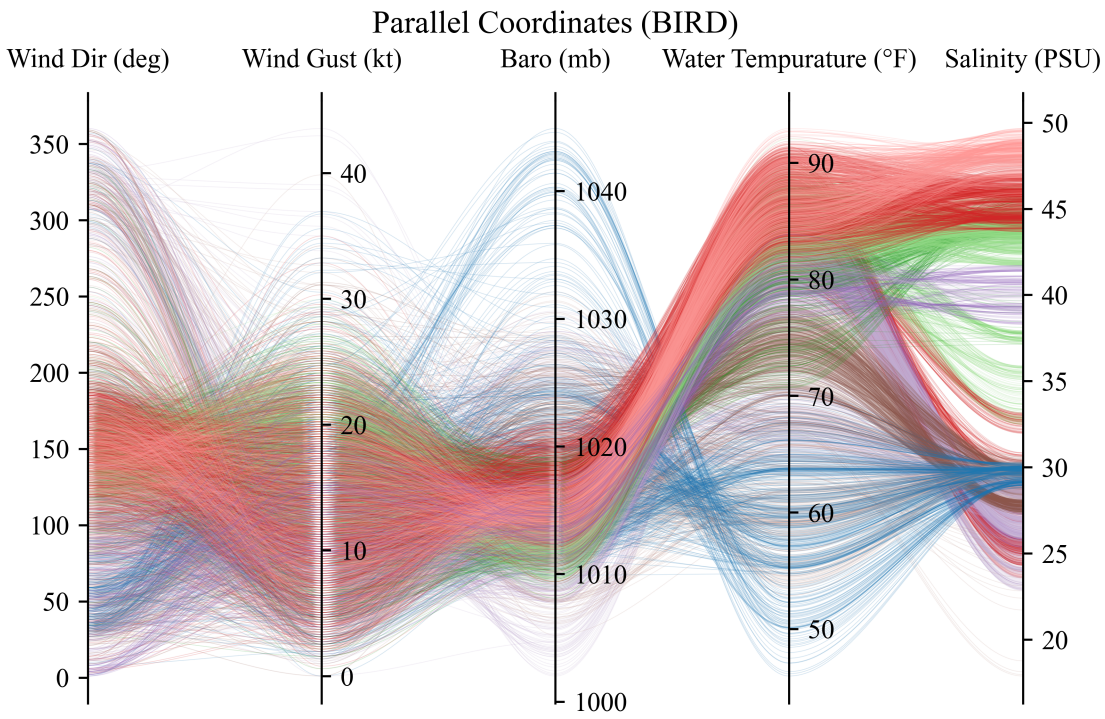

Figure 3.4. Parallel Coordinates plot of the TWDB's BIRD station data from 2007 - 2014, showing forecasts separated by month.

Salinity Data Analysis

The salinity data, as previously mentioned, comes from two data sources. These sources are the World Ocean Database [40] and the Texas Water Development Board's Water Data for Texas. Each of these sources have their own processing methods, as well as their own expected values. A great deal of the World Ocean Database's data used for this study covers the coast, past the Laguna Madre, while the data from Water Data for Texas covers areas within the Northern and Southern Laguna Madre areas. Furthermore, the data from the World Ocean Database contains samples of salinity at multiple depths, whereas samples from the TWDB are static on their gauge sations.

Furthermore, additional data was gathered from NOAA's gauge data, which happens to be nearby most of the TWDB stations. To gather a general idea of the typical values of the two gauge stations, and how they relate to one another, Figure 3.4 was created to show the relationship utilizing a parallel coordinates plot. Highlighted are the months of the year. There seems to be patterns

in the salinity, as well as a partial relationship to barometric pressure. With this is mind, we should see some positive results when adding this data to the model.

Given that the salinity data is gathered from different sources, there are differences in the amount of processing and quality checking done in the datasets. The data from the World Ocean Database undergoes automated quality checks, along with several optional processing methods that occur upon retrieval to transform the data towards the desired end product.

The salinity from the TWDB, however, is largely raw data from sensors. When retrieving the data, there is a reminder at each station that the data does not undergo quality checks, and that checks should be done by the user of the data. As such, the data was processed to drop some extreme outliers, as shown in Figure 3.5. While not a perfect method, the goal is to check whether or not forecasting is feasible. As such, some outliers such as this are acceptable.

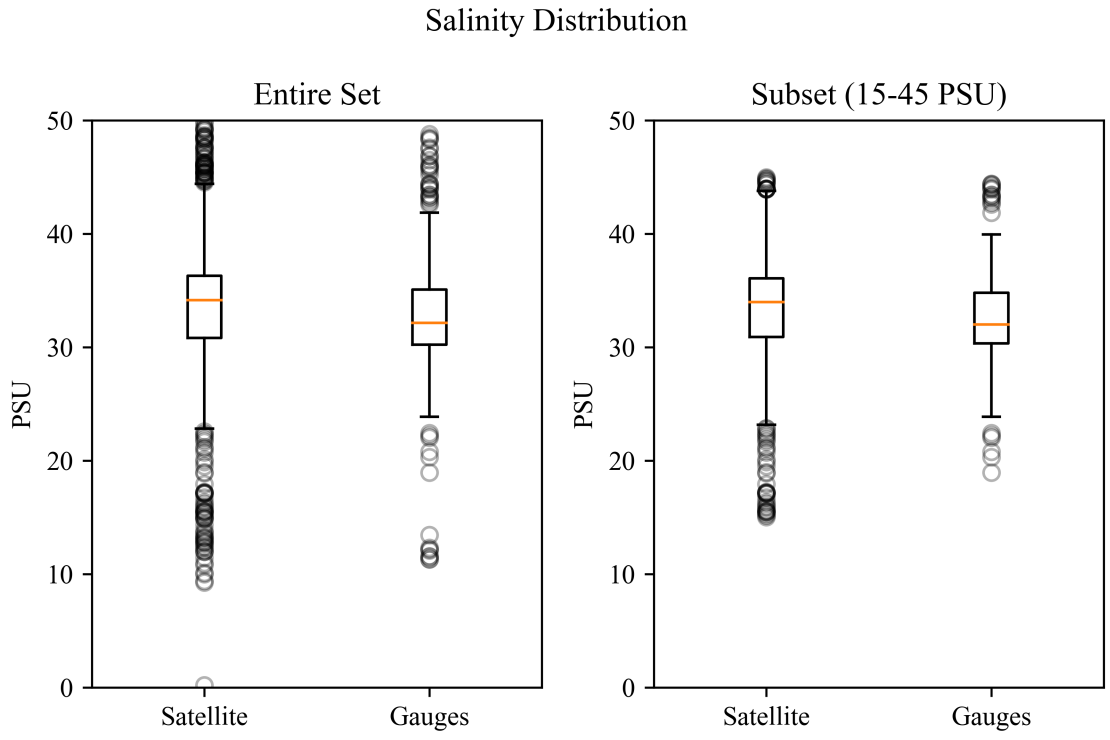

Figure 3.5. Box plot highlighting salinity distribution of datasets utilized.

Satellite Data Analysis

Analysis of the satellite data proved much more challenging, compared to the salinity data. This is given that there are multiple ways to try and tackle analysis of this data. Either by ignoring the spacial attribute, and analyzing the data as such, or by binning the data spatially for analysis as such. Each method provides its own upsides, as well as downsides.

By ignoring the spacial attribute, analysis of the satellite data becomes much simpler, however, localized trends are completely ignored and several data points could be incorrectly labeled as outliers and be considered for removal. A spatially aware analysis could provide much more information on possible local trends, however, there is the difficulty of how to partition the data particularly with respect to avoiding localized interactions being split between two points.

However, for the sake of simplicity, the existing Level-2 quality flags that are automatically generated for the data were used. The default quality flags for the Remote Sensing Reflectance were used, with both the good and questionable quality flags used for the Sea Surface Temperature. Given that the satellite data has already been preprocessed, hence the Level-2 naming of the data, the attributes should have had adequate preparation.

Data Extraction & Preprocessing

The processing of both the satellite and salinity data has undergone several changes throughout the process of improving the model's performance. Although the processes are similar, minor changes throughout the process were made in order to optimize the model's performance.

To begin with, the salinity data needs to be extracted from the Comma Separated Value (CSV)s retrieved from the TWDB Water Data for Texas (WDFT) website. After this information is extracted, the data is merged with its coordinate information from the gauges station's metadata. This will be used when merging the data with the satellite information.

Salinity data from the World Ocean Database was also utilized for broader salinity measurements, being off the coast rather than within the Laguna Madre itself. This data was retrieved

from NOAA's National Centers for Environmental Information (NCEI). The salinity data was then preprocessed into standard intervals utilizing the tools provided by the website itself.

Additionally, gauge station data from NOAA was utilized to test if the additional information provided would improve model performance. This data would first need to be retrieved from NOAA's Tides and Currents dataset. Once retrieved, they would then need to be mapped towards the closest salinity gauge station.

As for the satellite data, it is first retrieved from the Ocean Color Web dataset, provided by NOAA Earth Data. It is necessary to gather both SST and OC data in order to make accurate predictions.

These particular data points are called Level-2 data, as they are data that are processed after the satellite data has been received. After the data is retrieved, the data is then loaded by a program to be prepped for further processing. Once all the data has been retrieved and basic preprocessing of the data has been done, the full processing of the data may begin.

For creating a usable dataset from the previously mentioned sources, the two salinity sources would need to be merged first and foremost. Taking care to adjust for any timezone differences in the data, the salinity measurements can all be merged with each other. If using the extended gauge data from NOAA, then theWorld Ocean Database dataset will not be able to be utilized for the final set.

Prior to extracting key data from the satellite data, missing and invalid data must be handled appropriately. While multiple data filling techniques have been used, such as utilizing zeros to replace the missing data, the method that produced the best results in this particular case was using the global mean of the entire dataset. This data processing step was preformed exclusively on the SST and Rrs data.

Once salinity data has been merged, and invalid data in the satellite dataset is filled, work can begin on finding matching data within the satellite dataset. To do so, the date and time information provided by the salinity data is used to find a corresponding satellite image. Forecasts can be generated by applying the desired interval to the matching process.

As the images are found, they are tested to see if the quality of the data located around the sample point is to desired standards. The quality of both SST and Rrs are required to be of a certain standard, which for the most part, are using information provided by the dataset itself, i.e., the Level-2 quality flags.

Once all of these conditions are met, the satellite data is extracted in a 3x3 grid centered on the data point of interest, to then be stored into a dataset of valid sample points, containing the extracted grid of SST, Rrs, latitude, longitude, and optionally wind speed, gust, direction, and barometric pressure.

At the end of processing, the distribution of the salinity data is as shown in Figure 3.5, on the left-hand side. However, as is shown in the box plot, the dataset contains a number of outliers. Model performance was heavily affected by this. As such, a subset of the data was used to then train the model. The distribution of this subset is shown on the right-hand side of Figure 3.5.

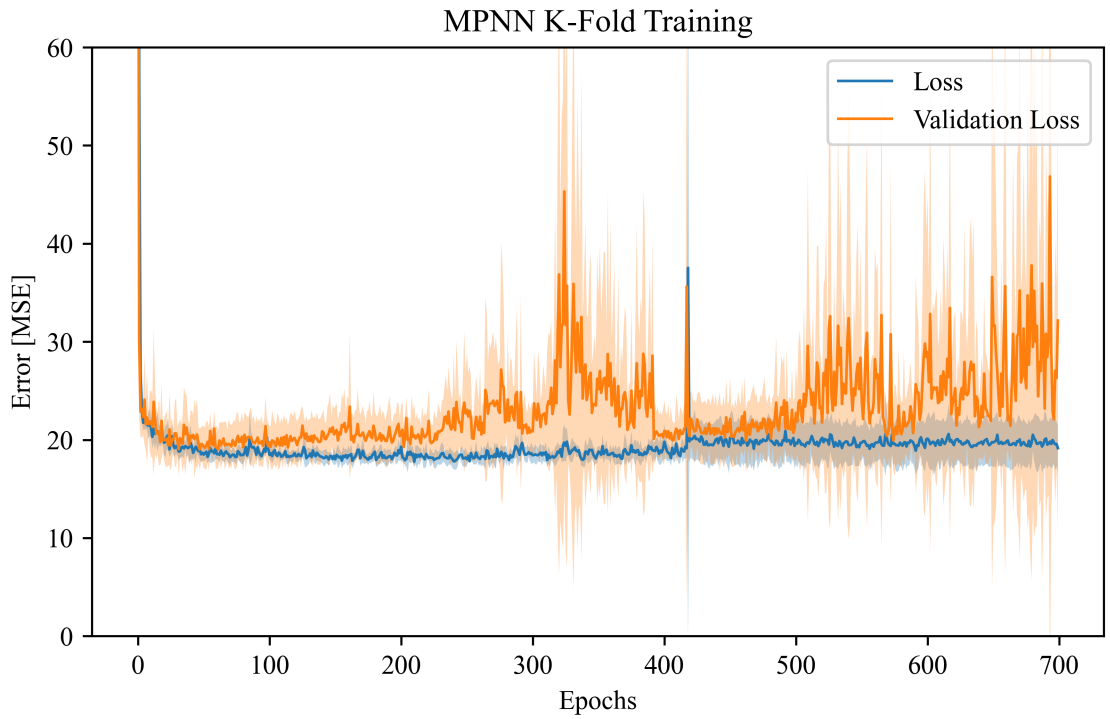

Figure 3.6. Mean loss of MPNN model with standard deviation shaded.

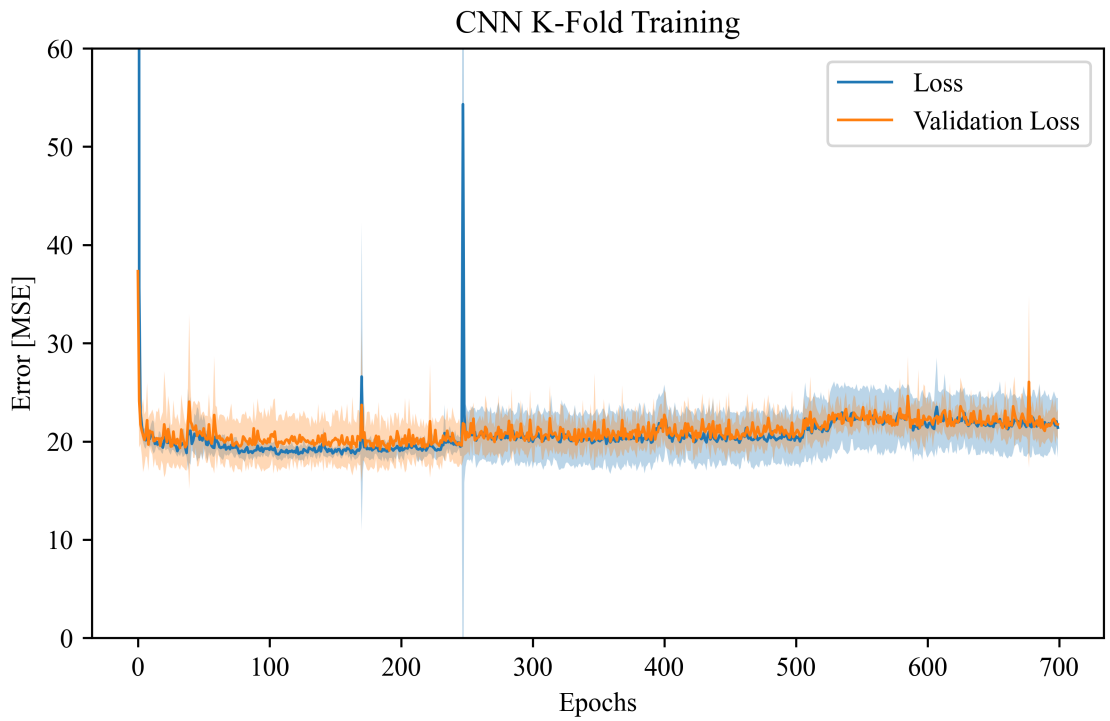

Figure 3.7. Training and test loss converge to roughly 20 MSE.

Modeling Process

The development of the models used to both predict and forecast SSS within the Laguna Madre initially started with a replication of the RNN model used in Song et al. [9], given the high degree of relevance to objectives of this paper.

However, after noticing the RNN model that was developed couldn't be adapted in its current state to utilize data from the MODIS-Aqua satellite, the MPNN model developed by Chen et al. [8] was chosen to be the starting point. From there, further architectural improvements were done using a random search.

The search parameters for the MPNN model are the number of layers hidden layers to be used, the number of neurons per hidden layer, whether or not the layers use batch normalization [43] and/or dropout layers [13], and the learning rate. The values used for the random search are shown in Table 3.1. Given the relative large search area for the number of neurons, a step size of 8 was chosen.

Additionally, the activation functions of relu, tanh, sigmoid, elu, and selu were tested to see which improved model results, along with starting to tune the batch normalization parameters, being momentum and epsilon. When evaluating the top 10 performing models again, no models using sigmoid activation were found and tanh seemed to be the most prevalent, with some relu in the mix. As such, sigmoid was dropped from further experimentation.

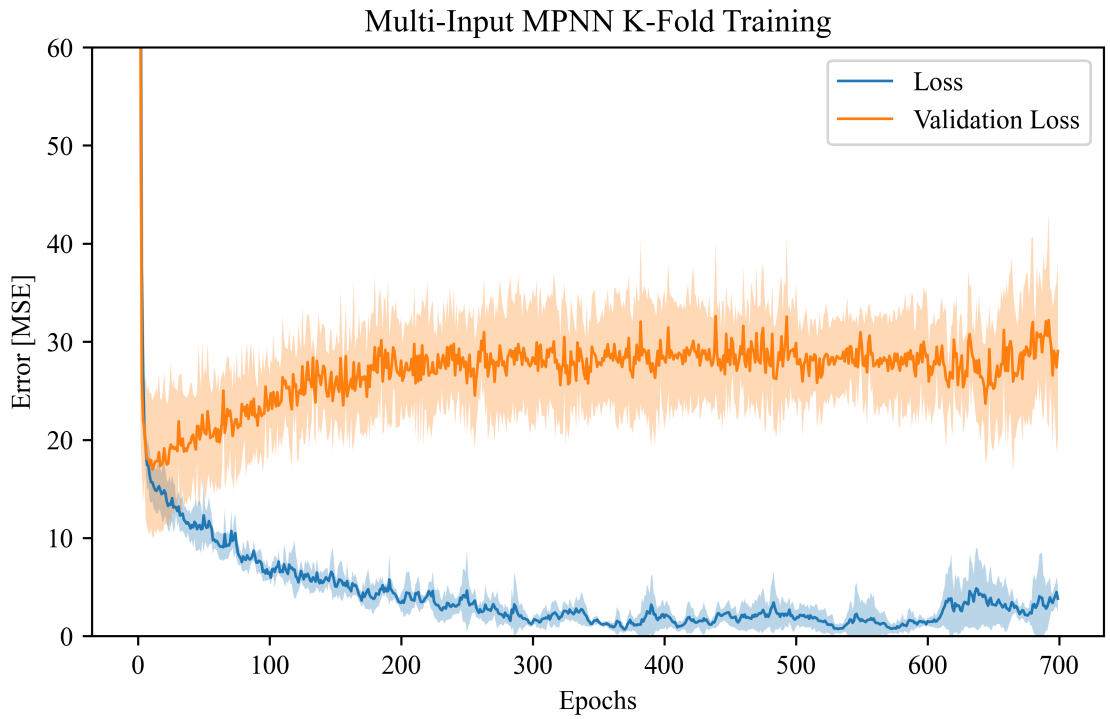

Figure 3.8. Multi-input MPNN loss converges to 0 MSE, unlike its satellite only counterpart.

After the initial tuning of the model, it was found that the model wasn't performing as expected. This was due to the fact that model optimizations were an iterative process, with assumptions that tuning one portion of the model and then tuning a separate portion would give optimal results. This sort of iterative approach was abandoned for a more general model architecture viewpoint. Additionally, given the relatively small size of the dataset, it was decided to forego the previous model and restart development of the same model using an alternate tuning technique.

The randomized search was then dropped for the much faster hyperband tuner [44], allowing for all parameters to be searched at once with little penalty to the overall runtime of the tuning session. This is due to the way that the hyperband tuner works. Several parameters were still

kept the same after some brief testing, such as the activation function being elu. Table 3.1 shows the hyperparameters used for tuning all of the models. It should be kept in mind that not every parameter is used for each model. For example, the number of CNN layers and filters isn't relevant to the MPNN based models.

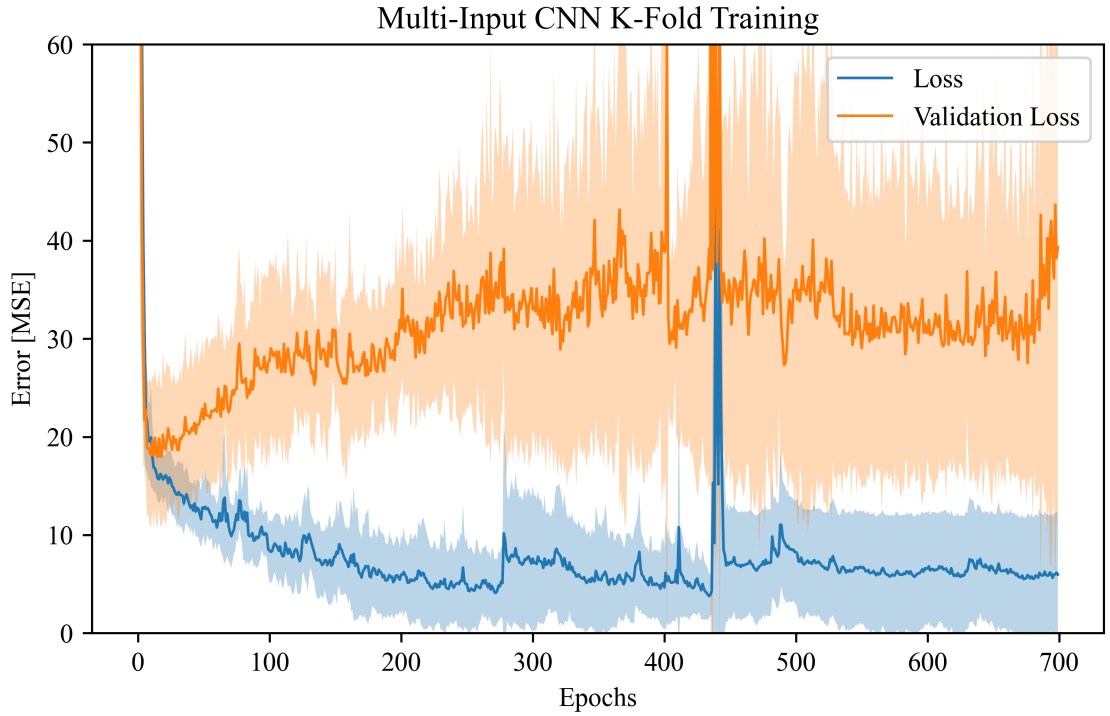

Figure 3.9. Multi-input CNN produces noisy loss, unlike the multi-input MPNN model.

The hyperband tuner [44] is a type of hyperparameter optimizer that is classfied as a Muli-fidelity optimizer. Hyperband is an extension of the simple successive-halving algorithm. By performing successive-halving with various configurations, and then adjusting the budget allocated to further explore the better performing models. However, it isn't without its faults. Hyperband doesn't care about its parameter spacing, meaning that it is more equivalent to Randomize Search than to Bayesian search.

The result of the hyperband based tuning for the MPNN model is show in Figure 3.6. Of note is the loss of the model isn't able to converge. This hints that either the data or the model itself isn't powerful enough to be able to solve the problem properly in this particular case. As such, a separate model was trained and then tuned to see if an alternate architecture could possibly converge.

Table 3.1. Architecture hyperparameters used for satellite-only hyperband based tuning.

Attribute	Values	
	Minimum	Maximum
Number of Convolution Layers	1	2
Number of Filters per Layer	1	5
Number of Hidden Layers	1	2
Hidden Units per Layer	8	128
Elu Alpha	-1	1
Learning Rate	0.005	0.09

Additionally, the hyperband based tuning method was used to then selectively tune the model using a CNN based architecture. The activation functions remained the same, however, the search space varied slightly, in that there was a variable amount of convolution layers, as well as a variable amount of kernels as is highlighted in Table 3.1.

Shown in Figure 3.7 is the result of tuning. Once again, it is shown that the model plateaus at around an Mean Squared Error (MSE) of 20. This could be due to a lack of information to be able to predict the Sea Surface Salinity accurately. After seeing the result of both of the satellite only based models, adding additional information via NOAA's gauges was done. This, however, reduced the size of the dataset from approximately 1000 data points down to 300. This also, lead to effectively removing the World Ocean Database SSS data. This was done to minimize errors due to the gauge's influence on the coastal salinity.

The models that take in the additional gauge data, which will now be termed the multi-input models, required the model's architecture to change in order to accommodate the two different data sources. For one of the inputs, the satellite input, it is a 3x3 grid of various satellite parameters stacked atop each other. However, the new input, the gauge input, is a single row of data, containing the coordinate information, along with the wind speed, gust, direction, and barometric pressure. This, as the name implies requires the model to have a sort of multi-input interface.

In this particular case, the satellite data and gauge data have essentially their own model that runs on them, with parameters similar to Table 3.1. The key difference being that the gauge data can actually have zero layers and pass by unprocessed. After the model has process the data separately,

Table 3.2. Architecture hyperparameters used for gauge dataset hyperband tuning.

Attribute	Values	
	Minimum	Maximum
Number of Satellite Layers	1	2
Number of Gauge Layers	0	2
Number of Combined Layers	1	2

the two data sources are combined and the process by one to two more layers of fully-connected hidden layers before being outputted. Table 3.2 highlights the additional parameters that can be tuned.

For further emphasis, the previously mentioned parameter table is also in conjunction with the previous Table 3.1. It can be though of as an extension to the previous parameters to allow for the processing of the added gauge data. Each parameter layer from Table 3.2 takes additional parameters from Table 3.1.

The final results of the model tuning are as shown in Table 3.3. The satellite only MPNN and CNN models have a higher degree of training loss due to the delicate balance of loss and validation loss.

When creating the final evaluations of the model, the learning rate was reduced on the plateau of model loss so allow for further convergence. The training and validation curves were utilized to determine the optimal stopping point based on the loss and validation loss.

Table 3.3. Results of model training on prediction dataset.

Model	Training Errors			Test Errors		
	MBE (PSU)	RMSE (PSU)	Index of Agreement	MBE (PSU)	RMSE (PSU)	Index of Agreement
MPNN	-0.49	4.03	0.660	-0.50	4.91	0.492
CNN	0.05	4.21	0.614	0.16	4.74	0.528
Multi-input MPNN	-0.06	3.02	0.794	0.00	3.44	0.603
Multi-input CNN	0.69	3.46	0.726	0.44	3.48	0.555

CHAPTER IV

RESULTS

Given the results shown in Table 3.3, only the output of the Multi-input MPNN model will be shown from this point on. The results following are based on the same model that was tuned using the prediction dataset, however, it was retrained using the one, three, and five-day forecast dataset.

The extraction principles for the forecast dataset are exactly the same process utilized for the prediction dataset, with the key difference being the time matching of the satellite image to the salinity data. The time matching for the forecasts are dependent on the desired intervals, as specified in the previous chapter.

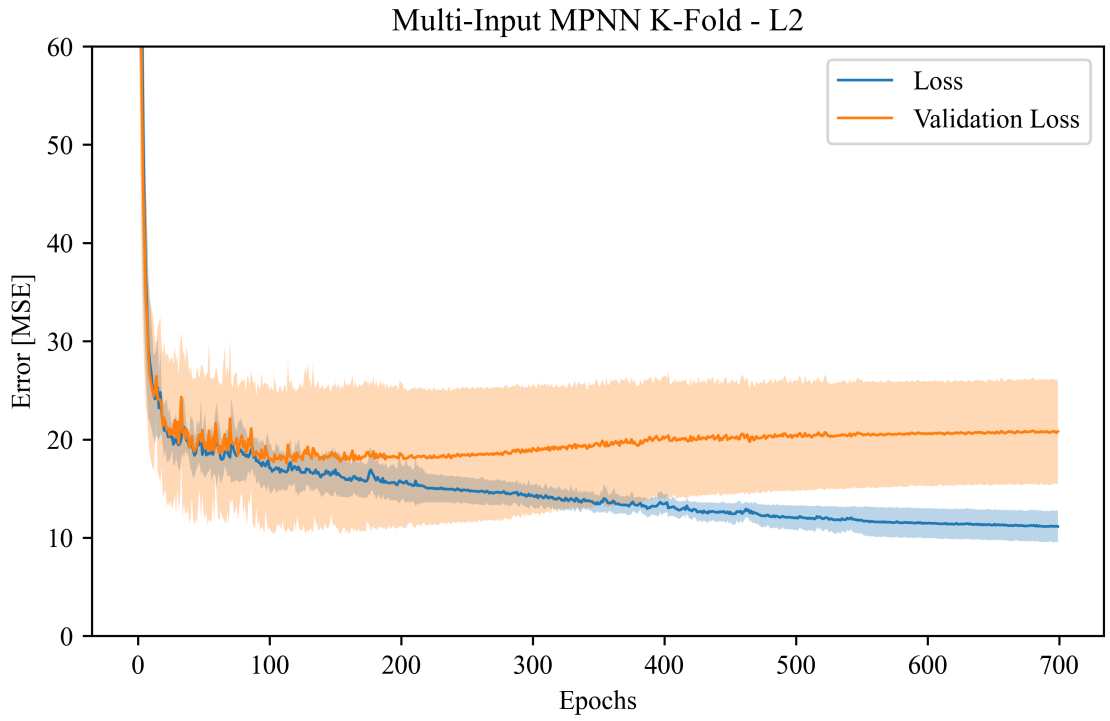

Figure 4.1. Multi-Input MPNN after L2 regularization and a dynamic learning rate.

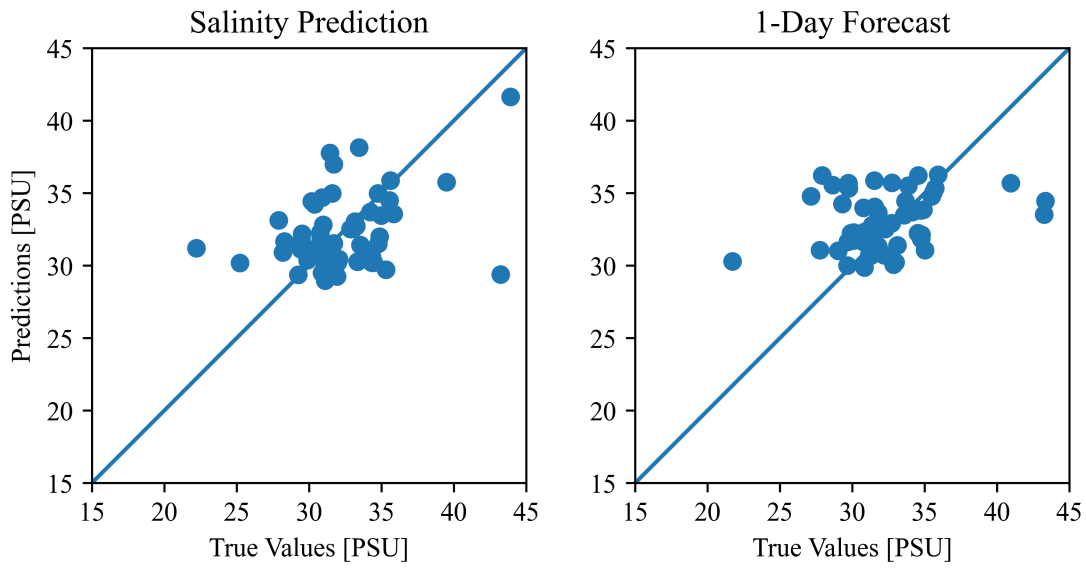

Figure 4.2. Model prediction and forecast on 0-day and 1-day test data.

First and foremost, the model training shown in Figure 3.8 highlights that the model is actually overfiting the dataset. This isn't a desirable result. As such, a further tuning was done to produce the same model with L2 regularization, who's training results are shown in Figure 4.1. The final parameter list for the model is shown in Table 4.1. Of note is that a decaying learning rate was used once the loss plateaued, to allow for further convergence than a static learning rate.

The model was then trained on the specified dataset, being prediction or one of the forecast datasets, using the training dataset to calibrate the model to the decided optimal training loss, and the evaluated using the test dataset. The datasets were randomly shuffled and split in an 80/20 manner.

Prediction Model

The prediction model is what estimates the current salinity give the input data, which would be satellite and gauge data in this instance. The left-hand image in Figure 4.2 shows the results of running the test data on the dataset. Of particular interest is the overall bias of the prediction model. Overall, the model seems to predict with no bias, meaning that only improvements to the accuracy are needed.

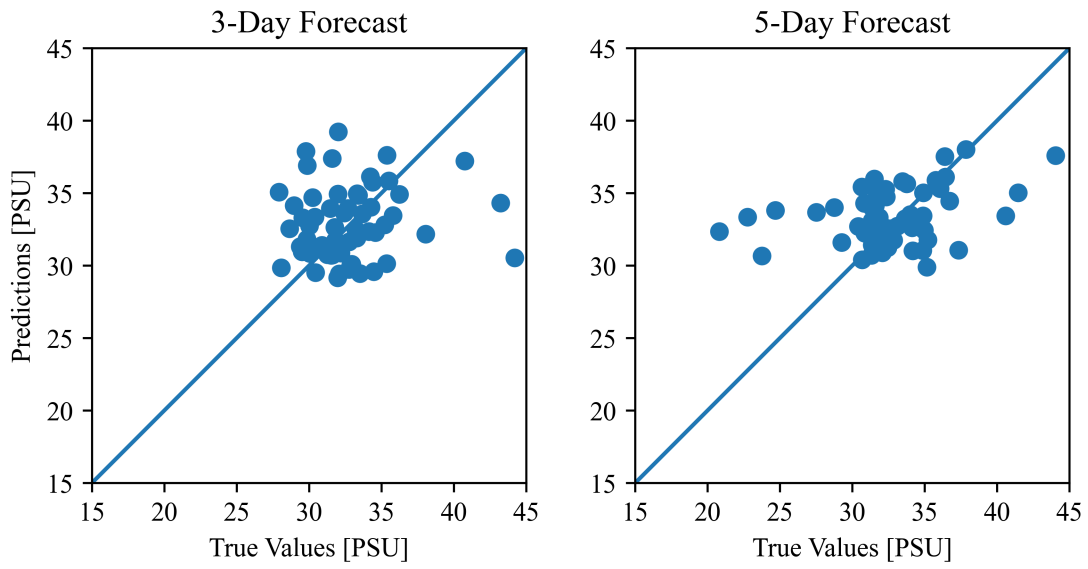

Figure 4.3. Model forecast on 3-day and 5-day test data.

The results of the prediction model show that it can relatively accurately predict the values that are roughly with in the first quartile through the third quartile range highlighted earlier in Figure 3.5. However, the model begins to have difficulty to predict salinity values that are outside the previously state range. This could be due to a few reasons, which will be discussed much more thoroughly in the following chapter.

This highlights, more than anything else, that to predict salinity in estuaries such as the Laguna Madre, more relevant parameters are needed for create accurate predictions. However, adding more parameters in area such as the Laguna Madre typically comes at the cost of having less data points to use overall.

Table 4.1. Multi-Input MPNN model hyperparameters.

Model Layer	Hidden Units	ELU Alpha	L2 Regularization
Satellite Layer 1	64	-0.35	0.432
Satellite Layer 2	24	-0.50	0.006
Gauge Layer	80	0.15	0.779
Combined Layer	60	-0.95	0.047

Forecast Model

As stated in the introduction to this chapter, the forecast models are simply the optimized prediction model that are then retrained on their relevant dataset. It should be noted that there are drawbacks to simply using the same model that are then retrained.

First and foremost, the downside to using the existing model instead of retraining the model leads to suboptimal performance, compared to a custom tuned model. However, there are still yet other methods to extract further performance from a tuned model by using an existing model. This discussion will be left for the following chapter.

The performance of the forecasting model shows that there a 1-day forecast accrues a small amount of loss increase, as in shown in Table 4.2. Figure 4.2 shows, on the right hand side, the cluster produced from the model output. Compared to the SSS prediction, the model clusters relatively similarly. The performance on the outliers, however, is slightly worse compared to that of the prediction model.

There is a further decrease in accuracy of the 3-day forecast, however, the 5-day forecast shows a slight improvement over the 3-day forecast. This small change may be of no significance and may be due to differences in the set distribution as well as the size of the respective forecast datasets.

Based off of both Table 4.2 and Figures 4.2 and 4.3, there is a clear decay in accuracy from an SSS prediction and 1-day forecast to a 3-day and 5-day forecast. This highlights the possibility for performing short-term forecasts with a standard deep learning model.

Table 4.2. Results of model training on forecast dataset.

	Training Errors			Test Errors		
Forecast Amount	MBE (PSU)	RMSE (PSU)	Index of Agreement	MBE (PSU)	RMSE (PSU)	Index of Agreement
Prediction (0-Day)	-0.06	3.02	0.794	0.00	3.44	0.603
1-Day	-0.03	3.48	0.671	0.72	3.51	0.519
3-Day	-0.11	3.18	0.800	0.06	3.80	0.433
5-Day	-0.04	3.56	0.654	0.61	3.76	0.518

CHAPTER V

DISCUSSION

While a number of different architectures were tried for the improvement of the forecasting model, such as adding a Neural Network to the model and adding several inputs further down the line to augment the model with stationary data points, there are several recent architectures that could be of note for future studies. One of these model types are Physics Informed Neural Networks.

Given that tidal and current data should impact how salinity is dispersed, there could be a study into effectiveness of modeling salinity using a Physics Informed Neural Network. This Type of architecture shows a great deal of promise, particularly in restricting the model into much more reasonable parameters based off of empirical modeling techniques.

Further improvements to the amount of detail the model produces can also be studied. There are several newer satellites orbiting and collecting the same, if not, similar data. The newer satellites generally have much higher resolution. As such, the model should be able to produce much finer details of salinity interactions.

Detailed information could potentially be extracted by utilizing the samples collected at multiple depths from the World Ocean Database dataset. The dataset itself contains depths of many meters, some reaching the thousands in deeper parts of the Gulf of Mexico. This, however, would prove difficult for the Laguna Madre as most sampling stations are set to a specific height with only one salinity sensor.

An additional improvement to the model could be utilizing forecast precipitation data, thereby aiding in producing forecasts. The method in which the precipitation could used may be modified

such that the precipitation can be in a gridded format. Conversely, it can be in a gauge station format. By running the forecasted precipitation through some modeling software, such as HEC-HMS, a rough discharge of fresh water in the Laguna Madre could be obtained, further augmenting the data inputs.

Given that DL models tend to tune better given more data, either expanding or selecting a larger time frame would be more appropriate. Given the cloudy nature of the area, a great deal of gauge station data-points were unable to be used for the training and evaluation of the model. A possible relief to this would be to utilize higher resolution satellite data, such as data from the Sentinel satellite mentioned earlier.

Additionally, future studies may attempt to use Level-1 satellite data, instead of preprocessed Level-2 data. This would be to utilize Deep Learning model's ability to extract information from the data, allowing for potentially more accurate data.

While a Recurrent Neural Network based model was not attempted due to difficulty adapting the data requirements from an existing model unto a new model, it should be possible to utilize satellite data directly to create an RNN model. Some Recurrent Neural Network models, such as neural ordinary differential equations, allow for data with continuous time. Additionally, other simpler methods for allowing the use of an RNN based model exist. An example of such a method would be adding a time delta into the input to allow for the model to accommodate for this automatically.

Further improvements could have been made to the forecasting models by developing a unique architecture for each of the forecasted times. This would have allowed for a the model to further adapt to tho problem at hand. While the problem of predciting and forecasting are similar, the problems are unique enough to warrant its own separate model.

Another possible method worth considering in future studies is the use of transfer learning. The Laguna Madre is not an area with plentiful amounts of data, however, other areas of the world have much more salinity sampling occurring. By training a model on their dataset, and then retraining the model on the data available in the Laguna Madre, the model performance may be better compared to that of a model solely trained in the local area.

Transfer learning isn't just exploiting Sea Surface Salinity information from separate locations, transfer learning can be preformed on a similar problem space in general. As such, utilizing transfer learning to then create the forecast models could see further benefits.

Given that estuaries are sensitive to both heat and water inflow, such that if conditions are right evaporation will cause the salinity levels to rise. Further data augmentation could be done to see if optically derived depth could be of use as an input parameter for the model.

While perhaps not plausible within the Laguna Madre, models should be able to predict and forecast salinity at multiple depths. The World Ocean Database dataset contains sample points at multiple depths, and can be standardized, allowing for the data salinity to be available at multiple depths for the purpose of training at standard intervals. This should allow for the tackling of stratification to a degree, however, the optical depth of the satellites may only allow for prediction up to a certain depth.

Each one of the previous improvements should allow for a much better performing model, however, advanced model architectures and modeling techniques should be done in areas with plentiful data to allow for much more room for experimentation. For areas with lesser amounts of data, transfer learning should be exploited if at all possible to allow for improved training without requiring as large a target dataset that would otherwise be required.

CHAPTER VI

CONCLUSION

Salinity is an important metric for monitoring the quality of estuaries around the world. This is particularly significant for the Laguna Madre, given its unique nature as one of the few hypersaline estuaries in the world. By monitoring salinity, the health of the estuary can be predicted and the current dynamics of the ecosystem predicted as well.

In order to produce the best forecasts for the area, multiple deep leaning models were created to determine the optimal model architecture and data processing approach to create accurate predictions.

Although the difference is relatively minor, the results suggest that the model would be able to produce forecasts of salinity for the Laguna Madre. At the very least, a one-day forecast should be feasible for the current modeling efforts of utilizing a MPNN based model.

While the results have shown that there is little change in the three and five-day forecast, more studies would need to be conducted to ensure that such extended forecasts are indeed feasable by utilizing model advanced Deep Learning based architectures.

As it currently stands, predicting and forecasting salinity in the Laguna Madre using remote sensing data, in combination with gauge data can be done. However, finer time resolutions would be much more difficult to produce without making the model biased towards the most prominent sources or using additional sources to further add more salinity data points.

While the temporal distribution of the model is lacking, the model itself could serve as a starting point for the augmentation of traditional existing models. This would help the existing modeling process, by allowing an initial starting salinity distribution from the deep learning model's results.

Of great hindrance to this model, as well as other forecasting models are the ability to predict salinity in areas where freshwater intrusion or more complex process occur. Further research can be done here to find appropriate data points that can closely model such processes to allow accurate forecasting. Further research can be done in areas with much more data to fully exploit powerful DL models.

In this particular case, the CNN based models performed on about the same level as the MPNN based models. This doesn't necessarily mean that the CNN based architecture wouldn't be useful for predicting and forecasting SSS. Rather, utilizing higher resolution data should allow for more opportunity for the CNN based model to perform better. Further research should be done on the use of CNN models in estuaries, and especially in the more open gulfs and oceans, where a great amount of quality satellite data may be.

Further studies can be done on producing a single model capable of providing multiple forecasts from one model, i.e. a one, three and five-day forecast from a single model. Depending on the source of the salinity data, as well as the general quality of satellite data in the area.

Overall, predicting and forecasting salinity in the Laguna Madre is challenging, given the relatively small amount of data compared to other areas. Further complicating the process is the satellite data available from the area. However, utilizing the MPNN based models, short-term forecasts of salinity in the Laguna Madre can be done.

REFERENCES

[1] Texas Parks and Wildlife Department. TPWD Bays: Upper Laguna Madre.

[2] N. Reul, S. A. Grodsky, M. Arias, J. Boutin, R. Catany, B. Chapron, F. D'Amico, E. Dinnat, C. Donlon, A. Fore, S. Fournier, S. Guimbard, A. Hasson, N. Kolodziejczyk, G. Lagerloef, T. Lee, D. M. Le Vine, E. Lindstrom, C. Maes, S. Mecklenburg, T. Meissner, E. Olmedo, R. Sabia, J. Tenerelli, C. Thouvenin-Masson, A. Turiel, J. L. Vergely, N. Vinogradova, F. Wentz, and S. Yueh. Sea surface salinity estimates from spaceborne L-band radiometers: An overview of the first decade of observation (2010–2019). *Remote Sensing of Environment*, 242:111769, June 2020.

[3] K. Fennel, R. Hetland, Y. Feng, and S. DiMarco. A coupled physical-biological model of the Northern Gulf of Mexico shelf: model description, validation and analysis of phytoplankton variability. *Biogeosciences*, 8(7):1881–1899, July 2011. Publisher: Copernicus GmbH.

[4] Huang I-Shuo, Lee J Pinnell, Jeffrey W Turner, Hussain Abdulla, Lauren Boyd, Eric W Linton, and Paul V Zimba. Preliminary Assessment of Microbial Community Structure of Wind-Tidal Flats in the Laguna Madre, Texas, USA. *Biology*, 9(8):183, 2020. Place: Basel Publisher: MDPI AG.

[5] Shuangling Chen, Chuanmin Hu, Brian B. Barnes, Rik Wanninkhof, Wei-Jun Cai, Leticia Barbero, and Denis Pierrot. A machine learning approach to estimate surface ocean pCO2 from satellite measurements. *Remote Sensing of Environment*, 228:203–226, July 2019.

[6] Zhiyi Fu, Linshu Hu, Zhende Chen, Feng Zhang, Zhou Shi, Bifeng Hu, Zhenhong Du, and Renyi Liu. Estimating spatial and temporal variation in ocean surface pCO2 in the Gulf of Mexico using remote sensing and machine learning techniques. *Science of The Total Environment*, 745:140965, November 2020.

[7] Saeed Rajabi-Kiasari and Mahdi Hasanlou. An efficient model for the prediction of SMAP sea surface salinity using machine learning approaches in the Persian Gulf. *International Journal of Remote Sensing*, 41(8):3221–3242, April 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01431161.2019.1701212.

[8] Shuangling Chen and Chuanmin Hu. Estimating sea surface salinity in the northern Gulf of Mexico from satellite ocean color measurements. *Remote Sensing of Environment*, 201:115–132, November 2017.

[9] Tao Song, Zihe Wang, Pengfei Xie, Nisheng Han, Jingyu Jiang, and Danya Xu. A Novel Dual Path Gated Recurrent Unit Model for Sea Surface Salinity Prediction. *Journal of Atmospheric & Oceanic Technology*, 37(2):317–325, February 2020. Publisher: American Meteorological Society.

[10] Tao Song, Jingyu Jiang, Wei Li, and Danya Xu. A Deep Learning Method with Merged LSTM Neural Networks for SSHA Prediction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–1, May 2020.

[11] Caimee Schoenbaechler, CG Guthrie, J Matsumoto, and Q Lu. TxBLEND Model Calibration and Validation for the Laguna Madre Estuary. *Austin, Texas: Texas Water Development Board, 60pp*, 2011.

[12] Ioannis Ioannou, Alexander Gilerson, Barry Gross, Fred Moshary, and Samir Ahmed. Neural network approach to retrieve the inherent optical properties of the ocean from observations of MODIS. *Applied Optics*, 50(19):3168–3186, July 2011. Publisher: Optical Society of America.

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, January 2014.

[14] S. C. Kleene. Representation of Events in Nerve Nets and Finite Automata. In *Representation of Events in Nerve Nets and Finite Automata*, pages 3–42. Princeton University Press, 1956.

[15] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, December 1989.

[16] C. J. Koblinsky, P. Hildebrand, D. LeVine, F. Pellerano, Y. Chao, W. Wilson, S. Yueh, and G. Lagerloef. Sea surface salinity from space: Science goals and measurement approach. *Radio Science*, 38(4), 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2001RS002584.

[17] Gary Lagerloef, F. Raul Colomb, Davide Le Vine, Frank Wentz, Simon Yueh, Christopher Ruf, Jonathan Lilly, John Gunn, Yi Chao, Annette deCharon, Gene Feldman, and Calvin Swift. The Aquarius/SAC-D Mission: Designed to Meet the Salinity Remote-Sensing Challenge. *Oceanography*, 21(1):68–81, March 2008.

[18] Jordi Font, Adriano Camps, Andrés Borges, Manuel Martín-Neira, Jacqueline Boutin, Nicolas Reul, Yann H. Kerr, Achim Hahne, and Susanne Mecklenburg. SMOS: The Challenging Sea Surface Salinity Measurement From Space. *Proceedings of the IEEE*, 98(5):649–665, May 2010. Conference Name: Proceedings of the IEEE.

[19] Yann H. Kerr, Philippe Waldteufel, Jean-Pierre Wigneron, Steven Delwart, François Cabot, Jacqueline Boutin, Maria-José Escorihuela, Jordi Font, Nicolas Reul, Claire Gruhier, Silvia Enache Juglea, Mark R. Drinkwater, Achim Hahne, Manuel Martín-Neira, and Susanne Mecklenburg. The SMOS Mission: New Tool for Monitoring Key Elements ofthe Global Water Cycle. *Proceedings of the IEEE*, 98(5):666–687, May 2010. Conference Name: Proceedings of the IEEE.

[20] Man Wong, Kwonho Lee, Joon Kim, Janet Nichol, Zhangqing Li, and Nick Emerson. Modeling of Suspended Solids and Sea Surface Salinity in Hong Kong using Aqua/MODIS Satellite Images. *–161– Korean Journal of Remote Sensing*, 23:161–169, January 2007.

[21] Sherry L. Palacios, Tawnya D. Peterson, and Raphael M. Kudela. Development of synthetic salinity from remote sensing for the Columbia River plume. *Journal of Geophysical Research: Oceans*, 114(C2), 2009. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2008JC004895.

[22] Jun Zhao, Marouane Temimi, and Hosni Ghedira. Remotely sensed sea surface salinity in the hyper-saline Arabian Gulf: Application to landsat 8 OLI data. *Estuarine, Coastal and Shelf Science*, 187:168–177, March 2017.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017.

[24] Maged Marghany and Mazlan Hashim. Retrieving seasonal sea surface salinity from MODIS satellite data using a Box-Jenkins algorithm. In *2011 IEEE International Geoscience and Remote Sensing Symposium*, pages 2017–2020, July 2011. ISSN: 2153-7003.

[25] Gary L. Powell, Junji Matsumoto, and David A. Brock. Methods for determining minimum freshwater inflow needs of Texas bays and estuaries. *Estuaries*, 25(6):1262–1274, December 2002.

[26] Erin A. Urquhart, Benjamin F. Zaitchik, Matthew J. Hoffman, Seth D. Guikema, and Erick F. Geiger. Remotely sensed estimates of surface salinity in the Chesapeake Bay: A statistical approach. *Remote Sensing of Environment*, 123:522–531, August 2012.

[27] Erick F. Geiger, Matthew D. Grossi, Arthur C. Trembanis, Josh T. Kohut, and Matthew J. Oliver. Satellite-derived coastal ocean and estuarine salinity in the Mid-Atlantic. *Continental Shelf Research*, 63:S235–S242, July 2013.

[28] Ryan A. Vandermeulen, Robert Arnone, Sherwin Ladner, and Paul Martinolich. Estimating sea surface salinity in coastal waters of the Gulf of Mexico using visible channels on SNPP-VIIRS. page 911109, Baltimore, Maryland, USA, May 2014.

[29] Anthony Vodacek, Neil V. Blough, Michael D. DeGrandpre, Michael D. DeGrandpre, and Robert K. Nelson. Seasonal variation of CDOM and DOC in the Middle Atlantic Bight: Terrestrial inputs and photooxidation. *Limnology and Oceanography*, 42(4):674–686, 1997. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.4319/lo.1997.42.4.0674.

[30] Chuanmin Hu, Frank E. Muller-Karger, Douglas C. Biggs, Kendall L. Carder, Bisman Nababan, Denis Nadeau, and Joe Vanderbloemen. Comparison of ship and satellite bio-optical measurements on the continental margin of the NE Gulf of Mexico. *International Journal of Remote Sensing*, 24(13):2597–2612, January 2003. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0143116031000067007.

[31] Paula Coble, Chuanmin Hu, Richard Gould, Grace Chang, and Michelle Wood. Colored Dissolved Organic Matter in the Coastal Ocean: An Optical Tool for Coastal Zone Environmental Assessment and Management. *Oceanography*, 17(2):50–59, June 2004.

[32] Rossana Del Vecchio and Neil V Blough. Spatial and seasonal distribution of chromophoric dissolved organic matter and dissolved organic carbon in the Middle Atlantic Bight. *Marine Chemistry*, 89(1):169–187, October 2004.

[33] Huasheng Hong, Jingyu Wu, Shaoling Shang, and Chuanmin Hu. Absorption and fluorescence of chromophoric dissolved organic matter in the Pearl River Estuary, South China. *Marine Chemistry*, 97(1):78–89, October 2005.

[34] Weidong Guo, Colin A. Stedmon, Yuchao Han, Fang Wu, Xiangxiang Yu, and Minghui Hu. The conservative and non-conservative behavior of chromophoric dissolved organic matter in Chinese estuarine waters. *Marine Chemistry*, 107(3):357–366, December 2007.

[35] D. G. Bowers and H. L. Brett. The relationship between CDOM and salinity in estuaries: An analytical and graphical solution. *Journal of Marine Systems*, 73(1):1–7, September 2008.

[36] N. V. Blough, O. C. Zafiriou, and J. Bonilla. Optical absorption spectra of waters from the Orinoco River outflow: Terrestrial input of colored organic matter to the Caribbean. *Journal of Geophysical Research: Oceans*, 98(C2):2271–2278, 1993. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/92JC02763.

[37] Y. H. Ahn, P. Shanmugam, J. E. Moon, and J. H. Ryu. Satellite remote sensing of a low-salinity water plume in the East China Sea. *Annales Geophysicae*, 26(7):2019–2035, July 2008. Publisher: Copernicus GmbH.

[38] Carlos E. Del Castillo and Richard L. Miller. On the use of ocean color remote sensing to measure the transport of dissolved organic carbon by the Mississippi River Plume. *Remote Sensing of Environment*, 112(3):836–844, March 2008.

[39] Steven E. Lohrenz, Wei-Jun Cai, Feizhou Chen, Xiaogang Chen, and Merritt Tuel. Seasonal variability in air-sea fluxes of CO2 in a river-influenced coastal margin. *Journal of Geophysical Research: Oceans*, 115(C10), 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2009JC005608.

[40] Tim P Boyer, Olga K Baranova, Carla Coleman, Hernan E Garcia, Alexandra Grodsky, Ricardo A Locarnini, Alexey V Mishonov, Christopher R Paver, James R Reagan, Dan Seidov, Igor V Smolyar, Katharine W Weathers, and Melissa M Zweng. WORLD OCEAN DATABASE 2018. page 207.

[41] NASA Ocean Biology Processing Group. MODIS-Aqua Level 2 Inherent Optical Properties Data Version R2018.0, 2017. type: dataset.

[42] NASA Ocean Biology Processing Group. MODIS-Aqua Level 2 Ocean Color Data Version R2018.0, 2017. type: dataset.

[43] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456. PMLR, June 2015. ISSN: 1938-7228.

[44] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

APPENDIX

APPENDIX

ACRONYMS

a_{CDOM} CDOM absorption coefficient. 15

ANN Artifical Neural Network. 8, 13, 14

BART Bayesian Additive Regression Trees. 13

BCART Bagged Categorical and Regression Trees. 13

CART Classification and Regression Trees. 13

CDOM Colored Dissolved Organic Matter. 15

CNN Convolutional Neural Network. vii, 2, 8, 9, 28, 29, 31, 40

CSV Comma Separated Value. 24

DL Deep Learning. iii, 2–5, 7–9, 11, 13, 15, 18, 37, 39, 40

GAM Generalized Additive Model. 13, 14

GLM Generalized Linear Model. 13, 14

GPU Graphics Processing Unit. 7

IOP Inherit Optical Properties. 18

MAE Mean Average Error. 31, 35

MARS Multivariate Adaptive Regression Spline. 13

MLR Multivariable Linear Regression. 14

MODIS Moderate Resolution Imaging Spectroradiometer. iii, 9, 10, 12, 13, 18, 19, 26

MPNN Multilayer Perceptron Neural Network. iii, vi, vii, 2, 8, 9, 11, 26, 28, 29, 31, 32, 34, 39, 40

MSE Mean Squared Error. vii, 27–29

NCEI National Centers for Environmental Information. 16, 24

NOAA National Oceanic and Atmospheric Administration. 16, 18, 21, 24, 29

OC Ocean Color. 5, 18, 24

OLI Operational Land Imager. 14

ppt Parts Per Thousand. 1

PSU Practical Salinity Unit. 6

REDOS Reanalysis dataset of the South China Sea. 10, 11

RF Random Forest. 13

RMSE Root Mean Square Error. 13, 31, 35

RNN Recurrent Neural Network. iii, 9–11, 25, 26, 37

Rrs Remote Sensing Reflectance. 6, 12, 14, 19, 23, 25

SeaDAS SeaWiFS Data Analysis System. 13

SeaWiFS Sea-veiwing Wide Field-of-view Sensor. 10

SSS Sea Surface Salinity. iii, vi, 1–3, 5, 6, 10, 12, 14–16, 25, 29, 30, 35, 37, 40

SST Sea Surface Temperature. 5, 6, 23–25

TPU Tensor Processing Unit. 7

TWDB Texas Water Development Board. vii, 10, 12, 17, 18, 20, 21, 24

WDFT Water Data for Texas. 18, 20, 21, 24

WOD World Ocean Database. 16, 17, 20, 21, 24, 29, 38

BIOGRAPHICAL SKETCH

Martin Jaime Flores Jr. attended the University of Texas Rio Grande Valley as an undergraduate in pursuit of a bachelors degree in Computer Science, and was awarded one in the winter of 2019. Shortly thereafter, Martin pursued and was awarded a Master's degree in Civil Engineering, with specialization of hydrology, in the Spring of 2022.

Martin has been awarded the Dwight David Eisenhower Transportation Fellowship in the years of 2021 and 2022, presenting his work at the 2022 South Texas All Hazards Conference. Additionally, he has presented his work on salinity forecasting at the 23 Annual Lower Rio Grande Valley Water Management & Quality Conference. Martin can be contacted via his email at martin.flores98@tutanota.com.