8-2024

Rapid Prediction of Buoyancy-Driven Exchange Flows at the Great Salt Lake: ML Models and a 1D Shallow Water Approach

Eric M. Larsen
Utah State University

RAPID PREDICTION OF BUOYANCY-DRIVEN EXCHANGE FLOWS AT THE GREAT SALT LAKE: ML MODELS AND A 1D SHALLOW WATER APPROACH

by

Eric M. Larsen

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Mechanical Engineering

Approved:

Som Dutta, Ph.D. Major Professor	Brian Crookston, Ph.D., P.E. Committee Member
Shah Muhammad Hamdi, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Vice Provost for Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2024

ABSTRACT

Rapid Prediction of Buoyancy-Driven Exchange Flows at the Great Salt Lake: ML Models and a 1D Shallow Water Approach

by

Eric M. Larsen, Master of Science

Utah State University, 2024

Major Professor: Som Dutta, Ph.D.
Department: Mechanical and Aerospace Engineering

In recent years, the use of data-driven models has significantly increased, influenced by advances in machine learning techniques, data availability, and the computational capabilities of modern hardware. In hydrology, with its wealth of measured data, machine learning has been used for flood pattern prediction, estimation of sediment loads, water quality, flow forecasting, and more. In the current thesis, we extend these methods to predict buoyancy-driven flows at the Great Salt Lake (GSL). The Great Salt Lake in Utah, USA, is a hyper-saline terminal lake separated into a northern and southern arm. The higher density of GSL's northern side and higher lake elevation of the southern side generate a buoyancy-driven exchange flow through the New Breach. Previously, flow through the breach has been modeled by Holley and Waddell [1] by numerically solving the 1D steady shallow water equation, and by Rasmussen et al. [2] using computational fluid dynamics (CFD). While the 1D model provides rapid prediction, it's accuracy regresses under certain lake conditions and the current implementation is not suitable for the geometry of the New Breach. The CFD model has high-fidelity, but the computational cost is unsuitable for rapid prediction.

To fill the aforementioned gap in predicting the buoyancy-driven exchange flow, this research leverages the availability of field data measured at the Great Salt Lake by United States Geological Survey (USGS), to develop models that predict the buoyancy-driven exchange flow between the north and south arms of the lake. The primary aim of this study is to quantify the suitability of data-driven models for predicting buoyancy-driven exchange flows. Multiple data-driven models have been developed and tested, including Linear Regression, Random Forest, Support Vector Regression, and Deep Neural Networks. The Fidelity of the machine learning models were compared against predictions from the physics-based models currently in use. The ML based approach has been found to be effective, but for lake and breach-geometry conditions that have substantial data to train models. Thus, in addition, a new formulation of the 1D steady shallow water model has been derived, improving on the model by Holley and Waddell [1]. The new derivation will account for trapezoidal cross-section and berm, present at the New Breach. The thesis shows that machine learning methods can accurately predict the complex exchange flow through the New Breach using lake conditions and measured velocity as input. It also illustrates the importance of 1D shallow water model, especially for conditions that lack the data to train the ML models, e.g. cases with different berm heights.

(115 pages)

PUBLIC ABSTRACT

Rapid Prediction of Buoyancy-Driven Exchange Flows at the Great Salt Lake: ML Models and a 1D Shallow Water Approach

Eric M. Larsen

The Great Salt Lake in Utah, USA, is a hypersaline terminal lake divided into northern and southern arms by the Union Pacific Railroad causeway since the 1950's. This separation has caused a difference in density and water surface elevation between lake arms. These differences result in a buoyancy-driven exchange flow occurring through an engineered breach in the causeway. Traditionally, modeling the flow through the breach has been done by numerically solving the 1D steady shallow water equations, and using computational fluid dynamics (CFD). The CFD models yield high accuracy results, but require substantial computing resources. This research proposes the use of data measured by United States Geological Survey (USGS) to create data-driven models to predict the exchange flow through the breach. The use of data-driven models, often referred to as machine learning, allows for faster flow prediction and requires lower computational cost compared to CFD simulations. This study uses, Linear Regression, Random Forests, Support Vector Regression, and Deep Neural Networks to create data-driven models from available USGS data. These models are compared to physics-based prediction models and monthly measurements taken by USGS. The results of this study show that data-driven models can accurately predict the buoyancy-driven exchange flow at a time consistent with USGS' sampling. These models could serve as a method for real-time prediction of the flow through the breach in the Great Salt Lake, facilitating better management of the flow between the arms of the lake and informing changes to the lake conditions over time.

To Elli and Amelia, for all their love and support. Reminding me to enjoy the little things.

ACKNOWLEDGMENTS

The chance I have been given to pursue a graduate degree will have a lasting impact on my life. I would like to thank my committee for their input and support. I am grateful to my advisor Dr. Som Dutta for his support and friendship throughout this process, continually teaching me new limits to what I can achieve. His dedication to exploring topics has taught me to look deeper and solve harder problems in steps. I want to give my appreciation to the Utah Water Research Lab for their help and support in this project as well as all the support from Utah's Department of Natural Resources. In specific I would like to thank Dr. Brian Crookston for allowing me to help with the field campaign for the Great Salt Lake. Most importantly I would like to acknowledge the love and support of my wife Elli. Thank you for patience and support throughout this experience. Finally, I would like to thank all those that have helped to support my research both directly and indirectly, your efforts have not gone unnoticed.

Eric M. Larsen

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

1D	One-Dimensional
ANN	Artificial Neural Network
ADCP	Acoustic Doppler Current Profiling
CC	Correlation Coefficient
CFD	Computational Fluid Dynamics
DNN	Deep Neural Network
DNR	Department of Natural Resources
GSL	Great Salt Lake
GABI	Generalized Area Based Index Model
HNN	Hybrid Neural Network
LR	Linear Regression
LSTM	Long-Short Term Memory Network
MSE	Mean Squared Error
ML	Machine Learning
NB	New Breach
NGVD	National Geodetic Vertical Datum
NSE	Nash-Sutcliffe Efficiency Number (related to R squared)
NS	North to South (related to discharge direction)
PBIAS	Percent Bias
PINN	Physics Informed Neural Network
RF	Random Forest
RNN	Recurrent Neural Network
SN	South to North (related to discharge direction)
SSWEM	Steady Shallow Water Exchange-flow Model
SVR	Support Vector Regression
USGS	United States Geological Survey
UPRR	Union Pacific Rail Road
UWRL	Utah Water Research Lab
WSE	Water Surface Elevation

CHAPTER 1

INTRODUCTION

1.1 The Great Salt Lake

The Great Salt Lake (GSL), located in northern Utah, is a hyper saline terminal lake, distinguished by its unique characteristics and separation into a northern and southern arm. This division was introduced in the late 1950s with the construction of the Union Pacific Rail Road's west-east running causeway [1]. The causeway, a rock fill structure incorporating box culverts, isolated the two sides of the lake while still allowing exchange flow. The main means of exchange flow between the arms was through the box culverts, with secondary exchange flow occurring by seepage through the causeway fill material.

The separation of the lake into north and south arms has a profound impact on their respective characteristics. The south arm receives 95 percent of the total freshwater for the lake from the Weber, Bear, and Jordan rivers, while the north arm relies solely on precipitation for its freshwater input [3]. This distinction results in a noticeable gradient in both density and water surface elevation(WSE) between the two arms. With recent observations revealing densities ranging from 1150-1200 kg/m^3 and 1050-1100 kg/m^3 in the north and south arm respectively. Throughout these density ranges the northern arm maintains a typical density 50-80 kg/m^3 higher than the south arm.

The difference in water salinity causes these two arms to act as separate ecosystems, with the southern arm being more biologically diverse. Due to these different conditions, the salt extraction industry is focused in the north arm, with the brine shrimp industry and local recreation being focused in the south arm.

Given the terminal nature of the lake, its water level is dependent on the precipitation received annually the upstream demands on the three freshwater sources-Weber, Bear, and Jordan rivers, and allowable exchange flow between arms. This dependency on external

factors influences the dynamic nature of the GSL and the need for a comprehensive understanding of lake dynamics to ensure proper ecosystem and lake water elevation maintenance.

One major contributor to lake conditions is the exchange flow between arms via the New Breach (NB) (See Fig. 1.1). Since the installation of the UPRR causeway, the GSL NB has been monitored by multiple agencies with different monitoring approaches [4], [5], [6]. These monitoring activities have lead to multiple data collection sites and methods. The monitoring schemes have ranged from meter measurements monitored on a 15-minute and monthly interval, studies to understand salt balance [7], to 1D models such as the model created by Holley and Waddell [1]. These monitoring efforts have been used to understand and predict the exchange flow through the causeway fill structure and the box culverts of the causeway.

While these methods have been able to help monitor the lake, there is still crucial information lacking in each of these methods based on the assumptions used. The major limitation of meter data is the locality of the data collected. While we can assume a general measurement is valid for the local area, it does not give a full depiction of the global conditions that exist at any given time. To understand full lake conditions would require taking measurements for the full lake; a field campaign with unrealistic expectations. Considering these limitations, attention is focused on the main modes of exchange flow with general lake conditions assumed from available monitoring stations around the NB.

An example of using the local data for flow prediction is the work done by Holley and Waddell [1]. In the work of Holley and Waddell, focus was placed on the box culverts of the causeway. Holley and Waddell's model used a method developed from three types of observed flow conditions, i.e. two-layer (bi-directional), arrested wedge, and single layer flow (uni-directional). Using the density of each arm, and their respective water elevations compared to the bottom of the culvert a discharge prediction could be produced. This model worked well for the box culvert flow and was implemented for the duration of the culverts service. With the semi-permanent nature of UPRR's causeway, there exists settling in the

fill structure each year. With increased settling the box culverts eventually were closed in 2013 [8] [3].

The closure of the culverts caused a reduction in the available flow between the south and north arms leading to a rise in the WSE of the southern arm. At its maximum, the difference in water surface elevation between the north and south arms was approximately 3 ft [9–11]. To increase the flow connectivity of the two arms the GSL NB was constructed December 2016. Instead of using box culvert, the NB instead is a trapezoidal breach cut through the causeway structure with a 150 ft bridge spanning the top allowing for an unobstructed channel flow (see Figure 1.1).

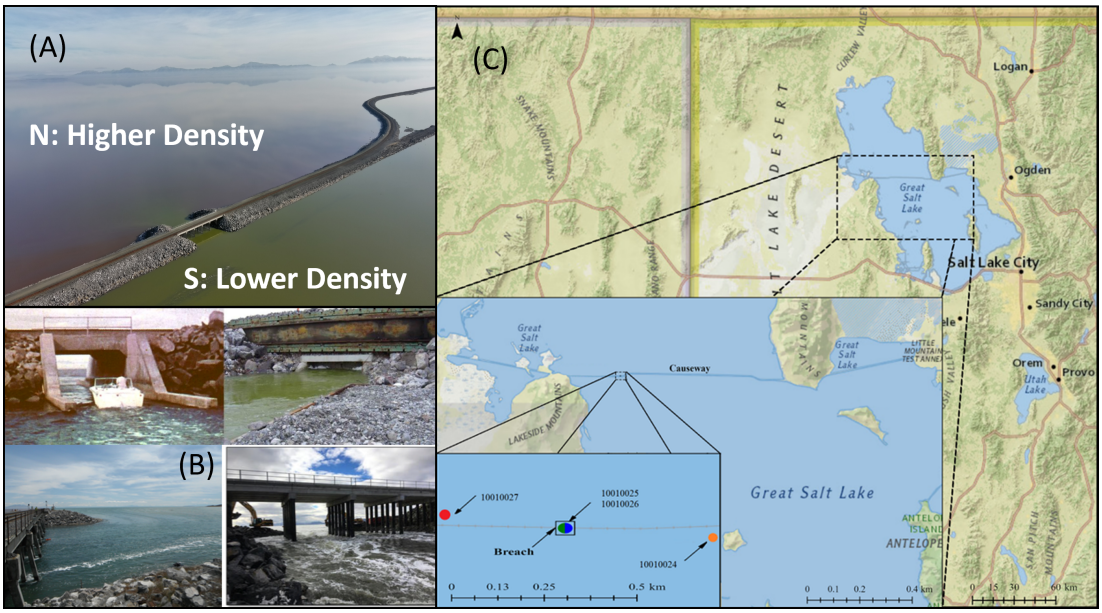

Fig. 1.1: Lake arm separation at the Great Salt Lake (A), New Breach structure (B), and location of the breach withing the Lake (C).

Over the next 6 months after the NB's opening, the GSL showed a decline in the difference between the north and south WSE measurements. Measurements of discharge performed during this period by the United States Geological Survey (USGS) showed a dominating flow from South to North (SN). Overtime this trend settled to a more stable bi-directional flow with two distinct flow layers. With the difference in density, ρ, and WSE

of the two arms, the expected discharge between the two arms is more complicated than a single height difference. Unlike standard water flow driven primarily by height difference, the water flow of the GSL also has buoyancy-driven effects. The density difference between arms allow the hydrostatic forces to create a buoyancy-driven flow from North to South (NS), while the SN flows are dominated by the height difference between lake arms, as shown in Fig. 1.2.

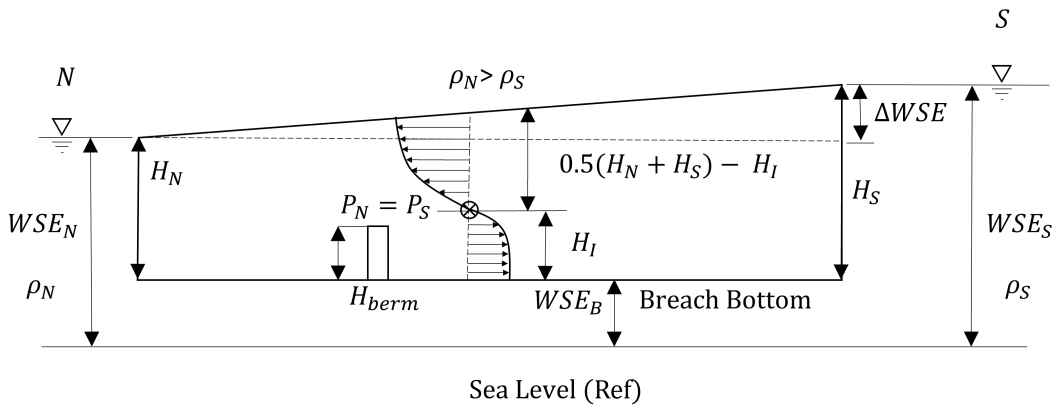

Fig. 1.2: Governing forces impacting the flow structure through the breach at the Great Salt Lake dictated by relative density and water surface elevation differences.

The competing factors of density and WSE cause a buoyancy-driven flow structure to exist at the NB. The primary mode of flow is a two-layer, or bi-directional flow. Additional flow cases do exist during more extreme lake conditions or during storm events at the GSL. In total, there are three primary flow cases for the GSL: 1) Bi-directional flow, where there are two layers of flow separated by an interface location. 2) Uni-directional South to North (USN), where there is purely south to North flow driven primarily by WSE difference between arms. Lastly, 3) Uni-directional North to South (UNS), where there is a single layer of flow NS due to density difference and correct lake conditions. Similar to the three flow regimes shown in work previously performed [1].

With the inclusion of the new NB there is a need for additional monitoring to understand flow in the new breach structure [4]. With the similarity of structures between

the closed culverts and the NB, the 1D model developed by Holley and Waddell has been adapted to predict SN and NS discharges for the NB. Due to improper adaptation of the equations to the different cross-section and channel roughness, the model is unable to predict the discharge with the same fidelity. To overcome the limitations of the adapted models for the NB, additional modeling was performed using computational fluid dynamics (CFD) given the known GSL geometry [12], [2]. The CFD model was created using the lake geometry localized around the NB and known lake conditions to solve for a quasi-steady state to extract discharge information given lake conditions.

Through a compilation of these simulation runs, the expected GSL discharge can be achieved compared to the USGS monthly measured discharge values. This CFD model has been fundamental in developing a rating curve to approximate discharge values given known conditions at the lake, and informing adaptive changes to the NB's control berm structure on the northern side of the channel. The limitation CFD simulations is the time required to achieve a steady state solution and the limited number of runs available. Typical CFD runs can take five to seven days to solve to steady state condition. This required time is non-conducive to a real time solving technique for lake dynamics. The GSL system is dynamic with lake conditions continually changing due to inflows, wind shear influences on WSE around the breach, and variations in density in each arm throughout the year. With these conditions in mind, one needs to have a model capable of adapting to these changing conditions while still being able to produce reliable results quickly. The availability of monitoring data around the NB lends itself to data-driven solving techniques like machine learning to predict exchange flows in the GSL.

1.2 Machine Learning in Hydrology

In recent years, data-driven models, especially machine learning (ML), have gained popularity both among the general public and within the scientific community. Machine learning takes on various forms depending on the application and data available. This study focuses on the use of Linear Regression (LR), Random Forest Regression (RF), Support

Vector Regression (SVR), and Deep Neural Networks (DNN) using measured data available around the GSL NB.

Linear regression is a fundamental statistical method used to produce prediction values given a number of inputs. Linear regression assumes that the relationship between the inputs and the desired output value is linear. The goal of LR is to find the best fit line for a set of observed data points, given a combinations of inputs. This fit is achieved by minimizing the difference between the observed values and the predicted values.

Random forest regression is a ML technique that uses a collection of decision trees to create a regression model capable of capturing non-linear relationships [13]. In RF regression, a group of decision trees is created, with each tree trained on a subset of data to learn patterns and relationships within the data. Each tree is developed by maximizing information gain as the number of leaf nodes is increased. The predictions of multiple trees are combined to produce a single prediction, leveraging the ensemble of predictions produced by individual trees.

The ensemble approach reduces the risk of overfitting and improves the robust nature of the generated model. While other methods can be sensitive to outliers, RF methods handle the effects of outliers by averaging across the forest. They are affected by outliers only in the trees that represent those data points. By averaging all outputs the effects are smoothed, resulting in a more stable prediction.

In hydrology, RFs are widely used for effectively modeling complex relationships between hydrological variables [14] [15]. Hydrological systems can be influenced by factors such as precipitation patterns, land use, soil properties, topography, and yearly trends. Random forests are capable of capturing these nonlinear relationships between predictors. One advantage of RFs compared to their machine learning methods is their interpretability. Where each tree is a collection of decision points based on variable values. Using this information RFs have the ability to assess variable importance among the predictors.

The ability for RFs to generate variable importance increases the focus on factors essential to prediction, while being able to assess the predictors less essential to accurate prediction. A major limitation of RFs though is their inability to create a governing equation given data point inputs. Instead of developing a prediction function, decision trees instead create an averaged prediction given the information found in the ensembled leaf nodes of the generated forest. This limitation requires the predictions to be within the dataset range to conform to expected trends. The are other methods capable of generating a governing function, methods such as LR, and SVR.

Support Vector Regression is another ML technique utilized in hydrology for flow prediction [16]. SVR is particularly effective in handling nonlinear relationships and high dimensional data, making it well-suited for modeling of complex flow processes like the GSL.

In SVR, the algorithm aims to find the optimal hyperplane that best represents the training data while maximizing the data explained by the margin, which is the distance between the hyperplane and the closest data points known as support vectors [17]. Unlike traditional regression methods that minimize the prediction errors, SVR focuses on minimizing deviations of predictions from a given ϵ-insensitive tube around the observed data points. The ϵ-insensitive tube allows for flexibility of the prediction around the hyperplane providing a robust approach to handle noisy data and outliers frequently observed in applications like hydrology.

In hydrology, SVR has been applied to areas including streamflow forecasting [14], rainfall-runoff modeling, drought prediction, and flood analysis [16]. Utilizing SVR, robust and accurate predictive models capable of handling outliers and noisy data can be generated. Given the intrinsic noise present in monitoring data collected from various instruments, the use of SVR offers a considerable advantage. Methods like SVR allow the developed method to accommodate these noisy data points either in the insensitive margin region or by excluding them from consideration in the produced hyperplane.

Artificial Neural Networks (ANN) and DNNs are predictive models used for a variety of purposes, ranging from image and text processing, numerical approximation, to data classification [18]. For each application, two primary considerations come into play: the nature of the collected data and the most suitable network for processing such data. Convolutional Neural Networks (CNN) are apt for handling image processing tasks [19], [20], [21], [22], while time series prediction benefits from network types like Recurrent Neural Networks (RNN) or Long-Short Term Memory (LSTM) networks [23], [24], where applications dominated by meter data, like the GSL, lend themselves to an ANN or DNN for relationship assessment [25] [26].

Each ANN/DNN is made up of a composition of layers, with each layer being a composition of nodes, and each node has a weight assigned to it, with each node in a layer sharing a bias vector for the layer. The layers of a network can be anything from the input information (input layer), the processing layers (hidden layers), or your output information (output layer).

Despite the differences in layers, there are two main kinds of models: regression and classification models. The latter of these model types is used to classify new information into discrete classes. Regression models, on the other hand, receive input data and output a continuous prediction using the model's developed relationships, where the relationships of each model are developed through the layers used in the network.

DNN layers can be of various types, including fully connected layers, convolutional layers, pooling layers, etc. [27]. The purpose of the layer structure is to map an M-dimensional input to an N-dimensional output [28] [19]. The weights and biases of the network link the M-dimensional input to the N-dimensional output, with the weights acting as the strength of one node's connection to another. The weights attached to each node have an influence on all the nodes in the following layers and are influenced by all nodes that precede it [27]. The state of each node is dictated by the prescribed activation function of the layer set. Activation functions can be ReLU, Tanh, Sigmoid, etc., where these activation functions are tailored to specific desired functionalities of the built DNN.

Neural networks are developed by training to a desired target given a loss function. Development is usually performed by dividing the total dataset into three sections: training data, validation data, and testing data. Training a network is done by allowing the network to create relationships between input and output parameters by tuning the weights and biases using a loss function. The loss function informs the network how well the DNN can predict the desired outcome, with a typical loss function being Mean Squared Error (MSE). The weights and biases are changed during training based on the learning rate and the gradient of the loss function to improve model performance [29, 30]. The amount of data the network processes at one time is called a batch, where any data partition smaller than the full training being considered a mini-batch. The determined batch size depends on the performance of a network and how clean the dataset is. Where larger batches are used for noisier data to mitigate the effect of noise. The weights in the network are updated each time the network runs through a batch of data through back propagation methods. When the network has looped through all batches in the training data it is called an epoch. Networks are trained for a set number of epochs or to a design performance condition. During training, the network is trained on the training set, and the performance is checked on the validation set. If the validation loss is consistently greater than your training loss, the network is trying to overfit to the training data [31].

Overfitting is when a model attempts to fit relations specifically to the training data and return those values to minimize the loss function. Having a model overfit to the training data is no longer useful for conditions outside the used training set, or that don't represent the training dataset well. There are ways to reduce the risk of overfitting by reducing the dimensionality of the model, tuning hyper-parameters, or by reducing the length a model is trained for. There is no direct answer to what hyper-parameters will result in the best data-driven model. Each model and dataset are individual and therefore must be tuned independently. The only way to know what model structure works best is by training a variety of model structures and comparing the results.

Every model has its drawbacks, and ML is no exception. However, its utility lies in domains with extensive datasets, with the increase in hydrological data collection in recent years [19], hydrology sets an ideal application for ML methods. In recent years, ML has found diverse applications within the hydrology community [19]. These applications yield themselves to ML due to the accessibility of data, and variety of data collected. Federal agencies like the United States Geological Survey (USGS), and other state agencies maintain historical measurements and data collection sights for various fields and measurements. Measurements ranging from streamflow measurements to satellite imagery of snow pack density and location.

The collected data from sources like this find further utility when used in data-driven models. Machine learning has used data like this to provide a wealth of applicable studies. Were ML as been used for flood pattern predictions in rivers [32], to estimate sediment loads in watersheds [33], assess water quality [34], make discharge predictions based on controllable gate parameters [28], conduct flow forecasting given temporal trend data [23], and forecast terrestrial water storage data given satellite imagery [20]. It is clear from studies as those by [19], [35], [36], [37], [38], the use of ML for hydrological and environmental applications continues to expand in both quantity and diversity.

In hydrological modeling, traditional methods rely on idealized physically-based numerical models to predict data. Instead, researchers can now use measurement data from real world systems to develop conclusions using ML. Machine learning though is not a cure all, and the system itself must be considered when deciding what modeling method is best suited for the specific application. For applications with sparse or noisy data ML is not well suited without considerable effort placed on data augmentation or further collection. In these cases a CFD or physically-based numerical model approach might be more appropriate for the given application. Nevertheless, hydrological applications like the GSL where data collection sites have been monitoring the lake for years lend themselves to ML models due to the quantity and quality of the data collected.

With the proven nature of ML for hydrological applications [19], ML showed promise in being able to assess the complex flow that exists in the GSL system. Where the GSL data is primarily meter data collected by USGS, using a LR, RF, SVR and DNN, it is hypothesized that a relationship can be created to determine the complex two-layer flow in the GSL NB using the data readily available through USGS monitoring stations around the NB.

CHAPTER 2

OBJECTIVES

The following research objectives are proposed for this thesis. Each objective is followed by research tasks required to accomplish the objective.

1. **Objective 1**: Quantify the efficacy of data-driven models for predicting buoyancy-driven exchange flows, in particular the flow through the breach in the causeway at GSL using lake conditions as the input.
    - Understand the physics of the flow, and study the measured data to finalize the lake-condition parameters relevant for flow prediction.
    - Compile available data from USGS collection sites located near the breach, and pre-process the data for developing machine learning models.
    - Develop models using Linear Regression, Random Forest, Support Vector Regression, and Deep Neural Networks for predicting flow through the breach.
    - Quantify the performance of the ML models against physics-based approaches, especially against 1D steady shallow-water equation.
2. **Objective 2**: Quantify the utility of using a hybrid approach (HNN), where prediction from 1D steady shallow water equation is used as additional input to Deep Neural Networks, for improving flow prediction.
    - Implement the 1D predictions as additional inputs to a NN, a parallel approach.
    - Implement the 1D predictions as the only input to a NN, a series approach.
3. **Objective 3**: Quantify the efficacy of using measured velocity as input to data-driven models, for predicting buoyancy-driven exchange flow through the breach at GSL.
    - Develop and test Linear Regression, Random Forest, Support Vector Regression, and Deep Neural Networks based models.

4. **Objective 4**: Understand the buoyancy-driven exchange flow through the breach in the presence of the new trapezoidal cross-section and berm structure; and develop physically-based numerical modeling techniques to predict the expected discharge behavior present at the New Breach.

    - Develop area based modeling technique using measured velocity and known channel cross section.
    - Re-derive the 1D steady shallow water equation based model to account for trapezoidal cross-section and berm influence.

CHAPTER 3

OBJECTIVES 1-3: ML MODELING OF GSL NB FLOWS

3.1 Methodology

The following chapter will cover the application of ML methods to predict the buoyancy-driven flow through the GSL NB structure. This chapter will explain the methodology for developing the ML models, including the data pre-processing, how these models are applied to the GSL problem and their accompanying results. For this study the developed ML methods are Linear Regression, Random Forest, Support Vector Regression, and Deep Neural Networks. Where each of these models will be built to perform the flow prediction task, though their approaches differ.Using these models one can assess how well the GSL system is suited for the use of data-driven models for flow prediction, and which modeling technique might be best suited for prediction of the NB exchange flows.

Linear regression explains the data using a linear equation fit using input parameters to make its prediction. Linear regression is a common hydrological tool used for making prediction. It must be understood that linear regression can be sensitive to outliers, but offers an understanding of what variables are important.

Random forest builds a series of decision trees to assess individual inputs compared to data the model is trained on. Using the comparison between points RF uses an averaged prediction over all trees to give a prediction. Random forest is an ensemble method lending itself to be more robust to outliers due to averaging.

Support Vector Regression works similar to linear regression, building a regression equation off of input parameters. The difference is in the consideration of the field, SVR uses a sensitivity range to assess what data points are considered, and anything outside the sensitivity range are not considered and deemed noise. This method allows for consideration

of noisy instrumentation consistent with real world sampling and can be robust to outliers in the dataset.

Finally, Deep Neural Networks model behavior by fitting a high dimensional function to the data points and then uses that to make predictions from an input vector. Utilizing DNNs can work for explaining high dimensional data and building a robust method to outliers. One major limitation of DNN is understanding what variables are important, where DNNs are more of a black box model capable of creating predictions; where their prediction process is much less interpretable compared to LR or RF methods.

3.1.1 Data Collection and Pre-processing

Regardless of what method is being used for ML prediction, one first needs to have an adequate dataset for modeling development. The dataset used for ML is just as important as the ML method used. More complex the data, more complex a model must be to approximate the patterns within the data. Complexity can arise from the number of instances in a dataset and the dimensionality of the collected data. Though, to understand the true complexity of a dataset, the data should be cleaned to eliminate spurious data that is not representative of the physical system.

For the GSL, there is no pre-existing dataset available for use. Instead, a dataset is compiled using the data available through USGS data collection sites, as seen in Fig. 1.1(C). The dataset compilation is performed by pulling data from the National Water Information System (NWIS) at data collection sites around the NB location. The localized data collection sites are sites: 10010024, 10010025, 10010026, 10010027 (see Fig.1.1C). The data collected from these sites has been provided on provisional status from USGS. These sites are then used to compile the following data at a 15-minute interval: Cell velocities 1-10, WSE for north, breach, and south locations, wind speed, and wind direction. The cell velocities are measured by USGS's Acoustic Doppler Current Profiling (ADCP) located on the northern side of the NB bridge. The ADCP discretizes the water column into 10 individual cells at set heights above the instrument blanking distance. More details The average velocity is recorded at each of these cells in the water column at a 15-minute

interval. An additional cell is added known as Cell 0 where the no-slip boundary condition is enforced for the bottom of the channel for the duration of the dataset. The monthly to bi-monthly information collected from site locations are density and specific conductance measurements with the addition of USGS measured discharge values for both NS and SN flows. The availability of data collected from the USGS data sites is shown below in Table 3.1.

Table 3.1: NWIS data availability for data collected in study.

Data Type	Time Interval	Date Begin	Date End
Velocity Cell 0	15 min	-	-
Velocity Cells 1-10	15 min	2017-09-13	2023-08-17
North WSE	15 min	2018-11-23	2023-08-17
Breach WSE	15 min	2017-09-13	2023-08-17
South WSE	15 min	2018-12-20	2023-08-17
Wind Speed	15 min	2018-05-17	2023-08-17
Wind Direction	15 min	2018-05-17	2022-12-12
Q SN (cfs)	15 min	2018-06-07	2020-06-04
Q NS (cfs)	15 min	2018-06-07	2020-06-04
Density North	Monthly	2017-09-13	2023-08-17
Density South	Monthly	2017-09-13	2023-08-17
SpCond North	Monthly	2017-09-13	2023-08-17
SpCond South	Monthly	2017-09-13	2023-08-17

Considering the difference in sample frequency and consistency of the different measurement systems at the data collection sites, there exist gaps in the data. Gaps in the data are caused by a variety of issues at the collection sites, either through the failure of the instruments to record data, or the instruments being locked at a fixed value etc.. To resolve the data gap issues, linear interpolation is used to fill in the small gaps existing in the dataset. Data cleaning is performed after to ensure fixed values over long periods of

time are removed from the dataset under the hypothesis of instrumentation malfunction for locations that have more variable data.

Linear interpolation is used to fill other gaps in the dataset with specific constraints in mind. Firstly, the interpolation is executed exclusively during periods of "steady" state conditions, characterized by minimal variation between the adjacent data points involved in the interpolation process. This criterion ensures the reliability of the interpolation results by avoiding instances of significant fluctuation. Secondly, the chosen time frame for interpolation is restricted to 30 minutes or less. This limitation is imposed due to the measurement frequency of USGS sites, which record data every 15 minutes. The rationale behind this decision is rooted in the understanding that shifts in the data do not occur at frequencies lower than this threshold. Lastly, particular attention is given to ensuring data points selected for interpolation do not exist within transition periods between the three flow cases under consideration in this study. This precautionary measure aims to prevent the introduction of inaccuracies that may arise during transitions and helps maintain the integrity of the interpolation in the context of the specific flow scenarios examined.

The use of these conditions provided reasonable assurance that linearly interpolated data points would produce results similar to expected conditions at the lake during the interpolated time frame. Any data segments that do not meet the interpolation criteria are excluded from the dataset. The use of interpolation enhances the information available to the model from the original dataset by filling the necessary gaps and removing noisy data where confident measurements or interpolation were not possible.

The dataset requires additional physical information to understand the density difference and its importance. The addition of density data has to be different from other measurements due to the sampling period intervals. Density and specific conductance measurements are measured on a monthly to bi-monthly basis. Therefore, the assumption that density remains relatively constant has to be made. This assumption was driven from monthly trends seen in lake conditions. In late 2020, the Utah Water Research Lab (UWRL) installed sondes in the north and south arms of the lake to measure the specific conductance

at a sampling rate of 15 minutes. These measurements are compared to the USGS water measurements. The resulting comparison can be seen in Fig. 3.1 and in Table 3.2. With the average specific conductance measurement from the UWRL and the monthly measurement from USGS being similar, the specific conductance measurement from USGS was used as the data for all entries of the measured month.

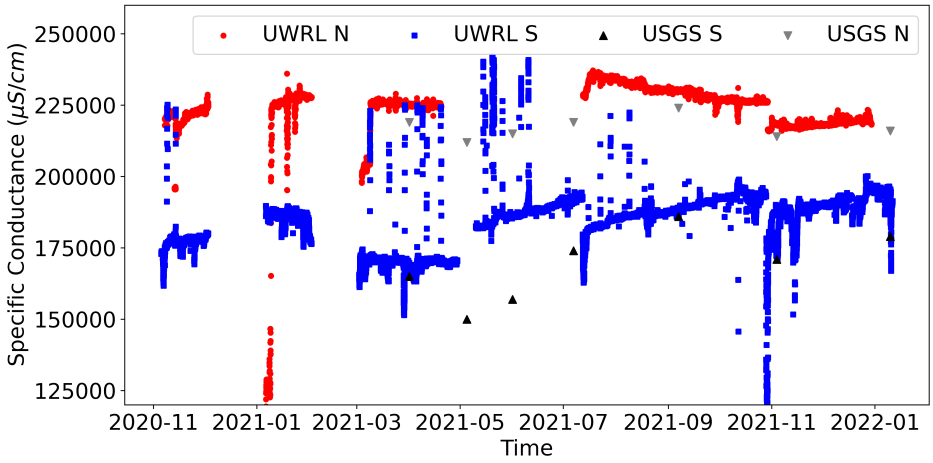

Fig. 3.1: Comparison of UWRL field measurement trend and USGS monthly measurement showing gradual transitions in lake arm conductance values.

Table 3.2: Specific conductance (μS/cm) comparison between UWRL and USGS monthly average measurements.

	South Arm			North Arm		
Month	USGS	UWRL	% Error	USGS	UWRL	% Error
Apr-21	165000	170398.92	3.27	219000	225050.10	2.76
May-21	150000	186159.04	24.11	212000	NA	NA
Jun-21	157000	188790.62	20.25	215000	NA	NA
Jul-21	174000	186606.34	7.25	219000	233986.90	6.84
Sep-21	186000	189331.75	1.79	224000	228788.47	2.14
Nov-21	171000	186827.49	9.26	214000	217806.06	1.78

Due to the comparison between the measured values, the monthly data from USGS was used as the measurement of specific conductance for all the measurements of that month in the data pool. The limited number of measurements from USGS restricted the number of data points available for the training of the model. After the construction of the breach in late 2016, HDR was required to sample the breach location for the course of three years. This information was taken on a monthly basis for the duration of the three years. Therefore, a comparison between the USGS and HDR specific conductance is assessed. The data from the HDR measurements are taken from the annual reports HDR provided to the state of Utah [4]. For the duration of the three years of sampling, two months are taken from each year and the measured specific conductance is compared, see Table 3.3.

Table 3.3: Comparison between USGS data and HDR annual report data for specific conductance (μS/cm).

	South Arm			North Arm		
Month	USGS	HDR	% Error	USGS	HDR	% Error
Aug-2017	145000	149600	3.17	218000	225000	3.21
Nov-2017	152000	149700	1.51	215000	219000	1.86
May-2018	138000	145300	5.29	215000	223000	3.72
July-2018	152000	149600	1.58	223000	223000	0.00
Feb-2019	166000	167100	0.66	222000	224000	0.90
Jun-2019	127000	137400	8.19	210000	225000	7.14

The results of the data comparison between HDR reports and USGS collection data shown in Table 3.3 show an acceptable difference between measurements, with the maximum difference being 8.19 percent. The comparison in measured values, HDR specific conductance data as the specific conductance information in the compiled dataset. The collected data has a more consistent behavior of sampling over the time frame of the dataset. The HDR reported data also collected density for the north and south arms over the time frame reported. Therefore, using these HDR reports, density and specific conductance data are added to the compiled dataset as a monthly average value given trends seen in UWRL's sonde data.

Compiling all the desired measurement data from USGS water data collection sites and from HDR annual reports results in a dataset consisting of 155,804 instances with 22 possible features. This dataset is split into two different datasets for prediction development, a hydrological dataset and a velocity dataset. The hydrological dataset contains WSE, wind, and density data with associated discharges. The velocity dataset contains cell velocities for cell0-cell10 and the associated discharges. Through refining the dataset and removing the measurement errors and instances with missing features, final datasets are created with 24,307 instances in the hydrological dataset, and 51,673 instances in the velocity dataset. Before the data is used for training, it is subdivided into training, validation, and testing datasets. This study uses a 80,10,10 split of the dataset for training, validation, and testing.

Inspecting the data pool, the percentage of uni-directional and bi-directional flow cases can be assessed. To characterize the uni-directional cases, the low-flow instances for the opposite flow direction are excluded. During high-flow events, the flow is known to be dominated by the flow direction. To understand this the data is standardized using the python library sklearn's standard scalar function [39] that standardizes the value given by the mean and standard deviation given Eq. 3.1.

$$Z_i = \frac{X_i - \mu_x}{\sigma_x} \tag{3.1}$$

Where Z is the standardized value, X is the variable of interest and σ is the standard deviation of the desired variable. Using this standardization the full dataset was standardized using the following values shown in Table 3.4.

Table 3.4: Standardization variables and associated values used for data scaling.

Variable	Mean (μ)	Std. Deviation (σ)
Cell1 [m/s]	-0.3458	0.3005
Cell2 [m/s]	-0.3763	0.3253
Cell3 [m/s]	-0.4222	0.3450
Cell4 [m/s]	-0.4630	0.3694
Cell5 [m/s]	-0.4325	0.3777
Cell6 [m/s]	-0.3080	0.3782
Cell7 [m/s]	-0.1105	0.3699
Cell8 [m/s]	0.1011	0.3632
Cell9 [m/s]	0.2862	0.3677
Cell10 [m/s]	0.3835	0.3724
WindSpeed [m/s]	2.2912	1.9504
WindDirection [θ]	184.2872	115.2449
Discharge S to N [m^3/s]	24.8718	17.2336
Discharge N to S [m^3/s]	14.7878	12.7054
NorthWSE [m]	1277.3427	0.4805
BreachWSE [m]	1277.6796	0.4649
SouthWSE [m]	1277.6661	0.4868
SouthSpCond [μS/cm]	160713.7209	18318.5945
NorthSpCond [μS/cm]	217928.7601	7126.303
SouthDensity [g/cm^3]	1.1000	0.0176
NorthDensity [g/cm^3]	1.2017	0.0279

Using this standardization the dataset is analyzed to asses the different flow cases that exist at the GSL, namely, bi-directional flow, and uni-directional flow both SN and NS. Plotting of scaled discharge for the total dataset is shown in Fig. 3.2

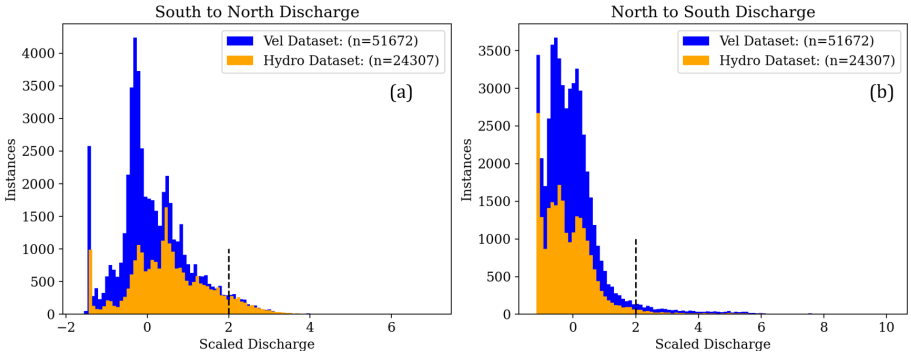

Fig. 3.2: Scaled discharge for GSL system for both SN and NS flows for cases classification.

From the results of Fig. 3.2 there is a clear Gaussian distribution for both discharge directions. With the majority of data consisting between a bound of 2 σ from the mean. From this assessment uni-directional flow cases were set to any flow existing above the 2σ threshold. Using this threshold bound the case separation of the dataset can be seen in Fig 3.3, with associated example velocity profiles for each of the cases.

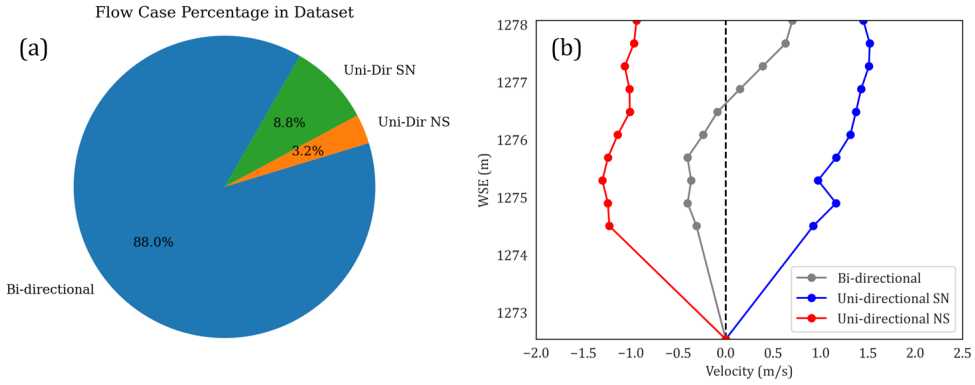

Fig. 3.3: Case separation in dataset (a) and example velocity profiles for each flow case (b) showing uni-directional NS flow (left) and SN flow(right), with majority bi-directional case (middle).

Here it can be seen that there is a class disparity between the bi-directional cases and uni-directional cases for the GSL dataset. This result is not unexpected due to the lakes usual nature to remain bi-directional with extreme flow cases arising under special circumstances like weather events.

3.1.2 Data Split for Model Development

In machine learning it is essential to understand and set limitations on what data is available to a model during training. If a model is able to see all data it has a tendency to overestimate the true model performance. In order to mitigate these issues the data can be divided into training, validation, and training datasets.

Each of these datasets represents an important aspect of ML development. Training data is the core data used for developing the model's predictive capabilities. This data section usually represents the largest portion of the datapool. In specific it is used to train the relationships between input and output variables. The model then checks performance using a validation set.

The validation set is used to verify the model is performing well for data similar to the training dataset, but is not the exact same. When there is a substantial difference between the training and validation accuracy there is a need for additional model development because performance is inconsistent. When models experience a higher validation loss this can be an indicator of the model overfitting to the training data.

Finally, the test section of a data split is used to evaluate the overall performance of the model. This allows one to compare the performance found in training to another dataset not found in the training. This final evaluation is used to assess how each of the models compare using the same set of data and determine relative performance.

For this study a randomized 80,10,10 training, validation, and testing split is used for model development. Where 80 percent of the data is used for training, 10 percent to validation, and the final 10 percent to testing. This is done to consider the overall size of the dataset. Since there is under 100,000 viable instances for training this study reserves most of the data for training, but leaves enough to have a well rounded evaluation of each

model's performance. In addition, the development of each model is done using a vectorized regression approach where each instance of the available dataset is considered independent of any other flow instance.

3.1.3 Performance Parameters

In any method employed to model a known phenomenon, it is crucial to establish performance metrics for result comparison. This study utilizes the Nash-Sutcliffe efficiency (NSE), root mean squared error (RMSE), correlation coefficient (CC), and percent bias (PBIAS) as performance metrics, as shown in Table 3.5. The NSE value is analogous to R^2 and is particularly relevant for discharge predictions.

Table 3.5: Performance metrics and formulas used for model evaluation.

Parameter	NSE	RMSE	CC	PBIAS
Equation	$1 - \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - \bar{Q})^2}$	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}$	$\frac{\sum_{i=1}^{n}(Q_i - \bar{Q})(\hat{Q}_i - \bar{\hat{Q}})}{\sqrt{\sum_{i=1}^{n}(Q_i - \bar{Q})^2 \sum_{i=1}^{n}(\hat{Q}_i - \bar{\hat{Q}})^2}}$	$\frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)}{\sum_{i=1}^{n} Q_i} \times 100\%$
Range	$-\infty \leq \text{NSE} \leq 1$	$0 \leq \text{RMSE} \leq \infty$	$-1 \leq \text{CC} \leq 1$	$-\infty \leq \%\text{Bias} \leq \infty$
Optimal Value	1	0	1	0

Note: Q_i = Observed $\hat{Q}_i$ = Predicted $\bar{Q}$ = Mean Observed $\bar{\hat{Q}}$ = Mean Predicted

By employing these metrics (Table 3.5), a comprehensive understanding of each network's performance relative to other models can be assessed. The primary metric in this research is Nash-Sutcliffe efficiency (NSE), indicating how effectively the model follows the predicted equals true value curve. The assessment considers all parameters, allowing an overall evaluation of the model's performance. As values approach the optimal values in Table 3.5, it indicates good model performance. The combined use of these metrics provides insights into how the model compensates for performance optimization without the need for complex techniques in ML visualization.

3.1.4 Machine Learning Models

For this study the following ML methods are developed to find a best performing method for the GSL NB system. In ML there is no one best answer to represent a given dataset. Any ML method must be assessed on the computation cost and the method's ability to model the patterns and relationships contained in the dataset. For this study four main ML methods are assessed; Linear Regression, Random Forest Regression, Support Vector Regression, and Deep Neural Networks. The evaluation of each developed model is done using the performance metrics of this study, including NSE, RMSE, CC, and PBIAS.

Linear Regression

Linear regression is one of the most fundamental statistical methods in hydrology used to produce prediction values given a number of inputs. The predicted values are created using an equation of form similar to Eq. 3.2.

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \ldots + a_nx_n \tag{3.2}$$

Here, y is the dependent variable, $a_1 - a_n$ are the weights associated with each independent variable, and $x_1 - x_n$ are the independent variables. The goal of LR is to find weights that minimize the residual sum of squares error between the expected and calculated values for the dataset. Linear regression for this study is implemented using Python package Scikit-learn (Sklearn) [39]. In the case of the GSL the available inputs are the systems WSE values for the south, breach, and north locations, the wind speed and direction data, and the respective densities of each arm of the lake. Using these values the produced. These inputs are used in multiple configurations to allow for LR regarding each of the water density information measurements individually for both north and south arms. Using these configurations two LR models are created, one model for each discharge both SN and NS:

1. Full input Specific Conductance: WSE North, South, North, Wind Speed, Wind Direction, Sp. Cond. S, Sp. Cond. N.

2. Full input Specific Conductance: WSE North, South, North, Wind Speed, Wind Direction, Density S, Density N.

While LR offers simplicity and interpretability to the dataset, it has built in assumptions that may not be appropriate for every dataset. Of these assumptions the primary assumption is linearity between the dependent and independent variables. Violations of this assumption can undermine the utility of LR to model a given behavior. For the GSL the flow is dependent on multiple parameters such as the water elevation difference and density difference between the lake arms. The non-linear response to change in variables may not be suitable for LR. Despite these limitations though, LR remains a valuable tool in hydrology due to versatility and ease of implementation. Due to this fact this method is used to serve as a baseline compared to other common hydrological practices.

Random Forest

In this study RFs are developed using the same inputs given to other ML models namely; WSE information, wind data, and density information for each arm of the lake using python's sklearn package [39]. Each of RFs is tested using both specific conductance or density as the density information to the system. The RFs are developed using a grid search method given different number of trees in the forest, maximum allowed depth and minimum number of samples per leaf node. This approach is utilized to find the optimal parameters focusing on the overall performance of the RF model. The developed models are assessed using the performance metrics used for all models in this study.

Support Vector Regression

Implementation of SVR for this study is done using the sklearn library in python [39]. The SVR is developed using two different kernel functions, the radial basis function (RBF)(Eq. 3.3) and Sigmoid kernel (Eq. 3.4). Using a grid search on these kernels and two training parameters, γ influencing the band of the kernel for RBF, and variable importance for Sigmoid and C for the trade off between margin and training error in the RBF kernel.

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2) \tag{3.3}$$

$$K(x_i, x_j) = tanh(\gamma x_i^T x_j + r) \tag{3.4}$$

Each of the SVR models produced are assessed using the selected performance parameters and compared to one another to find the optimal configuration.

Deep Neural Network

With the nature of the data collected for the GSL being meter data, this study developed DNNs to model the expected discharge dependent on GSL conditions. The approach taken during the development of this network was, first, to train multiple networks given different configurations to test for model accuracy on performance parameters. The developed method uses a mini-batch gradient descent method to train the networks; subdividing the training data dependent on batch size to constrain how many instances were seen during a given feed-forward step. Each model was tuned using variations in the number of layers (depth), the number of nodes per layer (width), the activation functions used on each layer, length of training, and batch size.

Specifically, the activation functions tested in this study are Sigmoid, ReLU, and Tanh. Weights for each model configuration are initialized using a random normal distribution. The Adam optimizer through the Keras TensorFlow library [40] is used for node weight and bias updates throughout training.

$$\sigma(z) = \max(0, z^{(k)})$$
$$z^{(k)} = \sum_{i=1}^{n}(x_i^{(k)} w_i^{(k)}) + b^{(k)}$$

Where σ is the activation function value, z is the node value, $x_i^{(K)}$ is the connected node value to the k^{th} node, $w_i^{(k)}$ is the associated weight between connected nodes, and $b^{(k)}$ is the bias value of the node k. For this study, Mean Squared Error is used as the loss function for model performance (Eq. 3.5). Other more complicated loss functions can be used, but MSE is the primary loss function used for robustness and ease of computation.

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2 \tag{3.5}$$

Networks that performed near the optimal value of these parameters are isolated for further development. Each network's training is monitored at each epoch to compare the relative performance between the training and validation loss. The desired curve during training is a smooth decline in the loss function for both training and validation loss where necessary. If there is erratic behavior in the training, i.e., when the validation loss varies above and below the training loss throughout training with no clear trend, the model is over-complicated for the modeled behavior, and the model needs refinement. Networks showing trends of overfitting are removed from possible model structures. The reduced models are then compared using the four performance parameters for evaluation. Models performing well are tested on the test dataset, and relative performance is then compared to training performance to check for overfitting of the networks to the training data.

3.1.5 Hybrid DNN

One of the major limitations to developed ML models is their dataset dependence. Meaning the relationships developed from input to output parameters is confined to the values available in the dataset. This limitation does not allow models to predict the desired behavior if the input parameters are outside the training bounds. When measurements are outside the dataset range an output value will still be given, but it does not follow the expected trend. In essence when input parameters are outside the bounds of trained values, the model attempts to produce the a desired output but it may not follow the modeled behavior trend. Instead the network follows the trained trends and returns values which can diverge from the truth. Therefore, DNN performance shown in training and testing can only be expected for values found inside their training range. This known issue with DNNs leads itself to the question of how can values outside the dataset compiled be used to anticipate future changes to the system.

The proposed use to increase DNNs ability to contribute useable results inside and outside the training value range is to combine DNNs with outputs with one dimensional output from the current GSL model used by Utah DNR. When combining a conventional DNN with a more physically-based model yields a Hybrid Neural Network (HNN). This combination will be done in two different configurations. One design will incorporate the 1D model in parallel while the other will combine it in series (see Fig. 3.4).

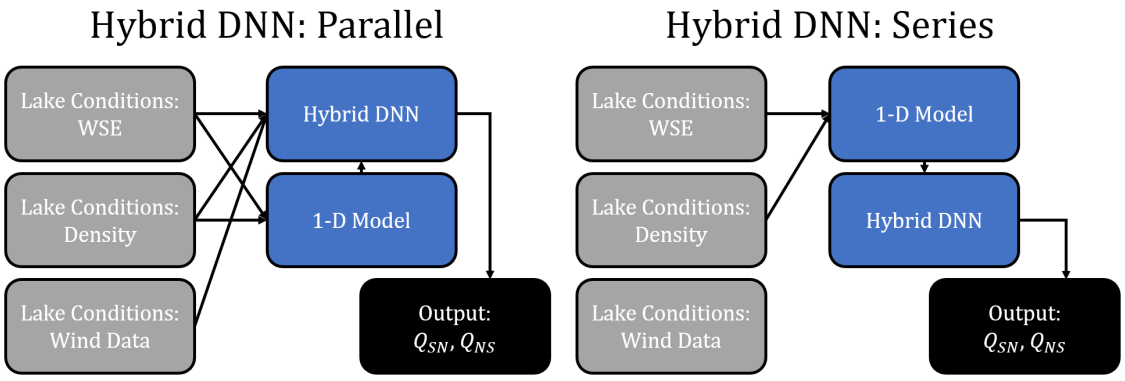

Fig. 3.4: Proposed structure for 1D integration for a Hybrid Neural Network method using parallel (left) and series (right) structures. The parallel method utilizing 1D output as additional input, and series using the 1D output as input to the HNN.

The parallel HNN is build off of the development of DNNs for GSL discharge prediction. This will be done by including the output from the 1D as two additional features to the input vector for the DNN structure. By doing so the system will be able to incorporate already known discharge values into the complex mapping procedure from the N dimensional input to the M dimensional output. This approach will leverage the total available data for lake conditions in the discharge prediction, and build off of current methods.

Another HNN is developed using the same 1D model but instead building the connection in series instead of parallel. In the series HNN the output from the 1D model is the only input to a DNN structure to refine and then output a desired discharge value. This approach will not include the wind data available from data collection sites due to lower data requirements for the 1D model. Here the only required measures are north and south WSE, and the densities of the respective sides of the lake. The DNN structure is build into the later end of the model to act as a refinement to the received discharge approximations given by the 1D model.

Using discharge predictions from the developed 1D model allows for increased boundary condition consideration. Unlike DNNs this 1D model is developed from fundamental principles based on concepts of conservation of mass and energy, seen previously in Section 3.6. These equations are not limited to a prescribed dataset, therefore, these models should

be able to increase DNNs ability to make accurate predictions outside the dataset. The 1D model was developed using fundamental assumptions leading to a more general solution. Assumptions such as uniform velocity for a cross-section both height and lengthwise, or the consideration that both layers are mutually exclusive with little interaction between layers. With idealized assumptions like these a loss of accuracy is expected from real world conditions, but this model has shown considerable versatility given the length of time it has been implemented for GSL discharge predictions.

While the equations used in the 1D model are not directly integrated into the training of the DNN, performance improvement is expected due to the preliminary discharge prediction produced by this model. Using these discharges it is expected that the DNN section will act as a refining process to the initial discharge prediction. The HNN will continue to implement the MSE loss function used previously (Eq. 3.5). This network's performance will be evaluated using the same performance parameters used for other models used.

3.1.6 Velocity Dataset Machine Learning

All former models developed in this study leverage the commonly used hydrology data to create data-driven models. This data being WSE, wind data, and density information. This data is key to many hydrological equations and flow predictive models. At the GSL there also exists another feature set collected in the dataset that explains the flow; the velocity profile data. If one knows the velocity data given a known cross-section a prediction can be made using simplified methods. These methods may not consider all available frictional losses that influence the expected discharge, but can serve as a good baseline approximation.

Given this knowledge, this study leverages the data from USGS's ADCP instrument positioned in the GSL NB channel. This data is more consistently available compared to all other data being available and allows for an increased dataset size compared to the hydrological based dataset. The velocity based dataset is of size 51,672 instances compared to the 24,307 instances of the hydrological dataset used to develop other data-driven models.

Using this increased dataset, the LR, RF, SVR, and DNN methods are developed to create flow predictions using the velocity dataset instead of the hydrological dataset.

3.2 Results

3.2.1 Hydrological Dataset Data-driven Models

Using the hydrological based dataset the following ML models are developed. Each method is the highest performing model given the grid search hyper-parameter tuning used on each of the models. Each of these models is developed for two different configurations, using either specific conductance or density as the density information for each lake arm. Each data-driven method produced two models and their relative performance is compared, with all performance parameters shown in Table 3.8.

3.2.2 Model Configurations

Linear Regression

The linear regression model is shown in (Eqs. 3.6-3.9) for the two model configurations. First the configuration for flow prediction using density is shown in Eq. 3.6 and Eq. 3.7.

$$Q_{SN} = 7.0240 \cdot H_S - 3.2255 \cdot H_B - 3.4204 \cdot H_N - 0.0330 \cdot V_{\text{wind}} + 0.0032 \cdot \theta + 0.0136 \cdot \rho_S - 0.0047 \cdot \rho_N \tag{3.6}$$

$$Q_{NS} = -6.0145 \cdot H_S + 1.9585 \cdot H_B + 3.6787 \cdot H_N + 0.2223 \cdot V_{\text{wind}} - 0.0364 \cdot \theta - 0.1195 \cdot \rho_S - 0.1156 \cdot \rho_N \tag{3.7}$$

Where Q is the expected discharge, H is the WSE height at a given location, V_{wind} is the wind velocity, θ is the wind flow direction from true north ($\theta = 0$), S is the specific conductance, and ρ is the density of a given side of the GSL. The equations for SN and NS flow using specific conductance instead are shown in Eq. 3.8 and Eq. 3.9.

$$Q_{SN} = 6.6932 \cdot H_S - 2.8447 \cdot H_B - 3.4701 \cdot H_N - 0.0273 \cdot V_{\text{wind}} + 0.0036 \cdot \theta - 0.0093 \cdot S_S + 0.1042 \cdot S_N \tag{3.8}$$

$$Q_{NS} = -6.2873 \cdot H_S + 2.3840 \cdot H_B + 3.7854 \cdot H_N + 0.2151 \cdot V_{\text{wind}} - 0.0342 \cdot \theta - 0.1538 \cdot S_S - 0.2148 \cdot S_N \tag{3.9}$$

Where S is specific conductance of each of the respective lake arms. From these equations it is clear to see that the relative importance of available variables is focused on the WSE variables compared to all others. Where the four later variables serve more as fine tuning of the overall prediction.

Using the results shown in Table 3.8, the use of LR is a viable means of predicting the exchange flow through the NB structure. The benefit of this method is its relative simple computation cost. The limitation of this model though is the assumption of a linear relationship between input and prediction variables. As variables violate this assumption the expected output will not reflect the true conditions expected at the GSL. For those a more complex method would be more appropriate.

Random Forest

Another method capable of assessing the importance of each of the variables are RF models. The developed RF model for flow prediction using specific conductance used a max depth of 20 using a minimum of 3 samples per leaf node, and a total of 500 decision trees. The model using density uses the same depth and minimum number of samples but instead used 1000 decision trees (see Table 3.6).

Table 3.6: Hydrologically based dataset RF model configurations.

Density/SpCond	Max Depth	Min Samples	N Trees
SpCond	20	3	500
Density	20	3	1000

The performance of the RF models is shown in Table 3.8 with the variable importance given these two configurations is shown below in Fig. 3.5.

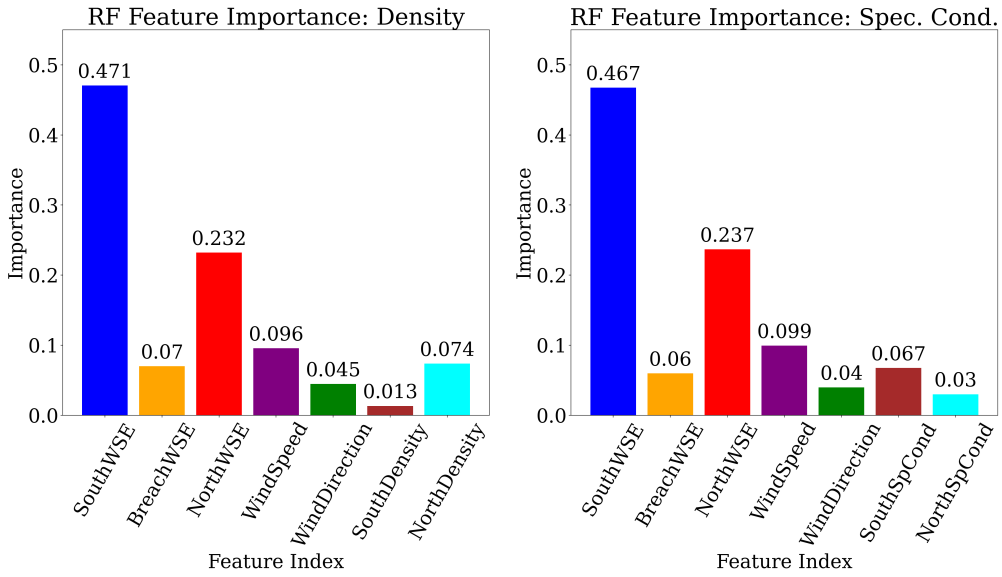

Fig. 3.5: Variable importance from random forest model using Hydrological dataset input.

The results shown in 3.5 demonstrate the importance of WSE for flow prediction. Compared to other parameters used in the RF method, WSE has an importance two to four times larger than other parameters. Using this importance this shows the WSE measurements used are the most important factors to consider when making a prediction of flow in the GSL. It is interesting to note that the breach WSE does not play as vital of a role compare to the other two WSE values. From a physical analysis of the system this finding can make sense. The breach WSE is measured at the bridge location in the channel (see Fig. 1.1) essentially the middle ground between the two lake arms. Using the south and north WSE and wind shearing effects one could estimate this parameter. Due to this fact one could see this as a derived parameter instead of an independent parameter.

In addition to this finding it is interesting to see the differences between the two model's variable importance. When using density the secondary contributing factors after WSE are wind speed, direction, and the northern density. Here the southern density has little

importance in the final decision made by the RF model. In contrast, when using specific conductance in the RF model all variables become secondary contributors with similar scale of their values compared to the major contributors, namely north and south WSE.

Support Vector Regression

For SVR models in this study this method needs to be broken into four different models like LR. With one model for each flow direction, and density information configuration. The resulting four model configurations are listed in 3.7. Compared to other methods that can formulate multiple equations to represent the mapping from input to output variables, SVR creates a equation similar to LR. Each equation has it's own ϵ insensitive region best suited for it's prediction, in this study the insensitive region is given based on the γ value given to the kernel function.

Table 3.7: Configurations for Support Vector Regression using hydrological dataset.

Flow Dir.	Density/SpCond	kernel	C	γ
SN	Density	RBF	100	1
NS	Density	RBF	100	0.1
SN	SpCond	RBF	100	1
NS	SpCond	RBF	10	1

Here since all configurations use the RBF kernel the γ value is inversely proportional to the bandwidth of the Gaussian kernel. From this result it can be seen that the NS Density configuration requires a larger bandwidth to allow for improved performance. This pattern is not unusual for GSL flow prediction as the prediction of NS flow has a higher uncertainty and thus the harder problem as shown in the results of Table 3.8.

Deep Neural Networks

Given the grid search method testing different number of hidden layers, activation functions, and layer widths (number of nodes) the best performing DNN in this study had a [100,100,100,100] configuration meaning it had four hidden layers with 100 nodes in each of the layers. Given the input size of 7 total inputs, and 2 outputs this results in a total of 31,300 trainable parameters, with the best performing activation function being ReLU.

3.2.3 Model Performances

With each model having its own strengths and weakness it is necessary to compare their performances given the same data. In Table 3.8 each model configurations performance given the tested parameters is shown compared to one another.

Table 3.8: Statistical values of data-driven models using hydrological dataset.

Model Type	Flow Dir.	Density/SpCond	NSE	RMSE	CC	PBIAS
LR	SN	Density	0.7638	0.4860	0.8741	2.4018
	NS	Density	0.6838	0.5623	0.8277	22.8270
	SN	SpCond	0.7669	0.4828	0.8759	2.2797
	NS	SpCond	0.6897	0.5571	0.8313	21.7333
RF	SN	Density	**0.8553**	**0.3804**	**0.9250**	**1.4067**
	NS	Density	**0.8181**	**0.4265**	**0.9047**	**4.8086**
	SN	SpCond	**0.8565**	**0.3788**	**0.9257**	**1.2454**
	NS	SpCond	**0.8187**	**0.4258**	**0.9051**	**5.4812**
SVR	SN	Density	0.8146	0.4306	0.9026	-1.7867
	NS	Density	0.7971	0.4504	0.8932	-21.8328
	SN	SpCond	0.8128	0.4327	0.9016	-1.7047
	NS	SpCond	0.7859	0.4627	0.8873	-17.7138
DNN	SN	Density	0.8081	0.4381	0.8992	3.6342
	NS	Density	0.7914	0.4567	0.8898	-15.3282
	SN	SpCond	0.8090	0.4370	0.8997	-3.7007
	NS	SpCond	0.7840	0.4647	0.8860	-13.2196

From these results it is clear to see that RFs are the best performing method for flow approximation using the hydrological dataset. Each of the methods besides LR perform relatively similar to RF irrespective of the increase in complexity in the SVR and DNN methods. Here RF methods are capable of performing well with a simplified means of approximation, comparing known flow values to desired values using numerical decisions based on input parameters. Though evaluation can be done purely numerically, these models are assessed visually as well to highlight where predictions have issues and possible reasons for those issues. These results are shown below in Fig. 3.6 and Fig. 3.7 showing both configuration sets of models based on density information.

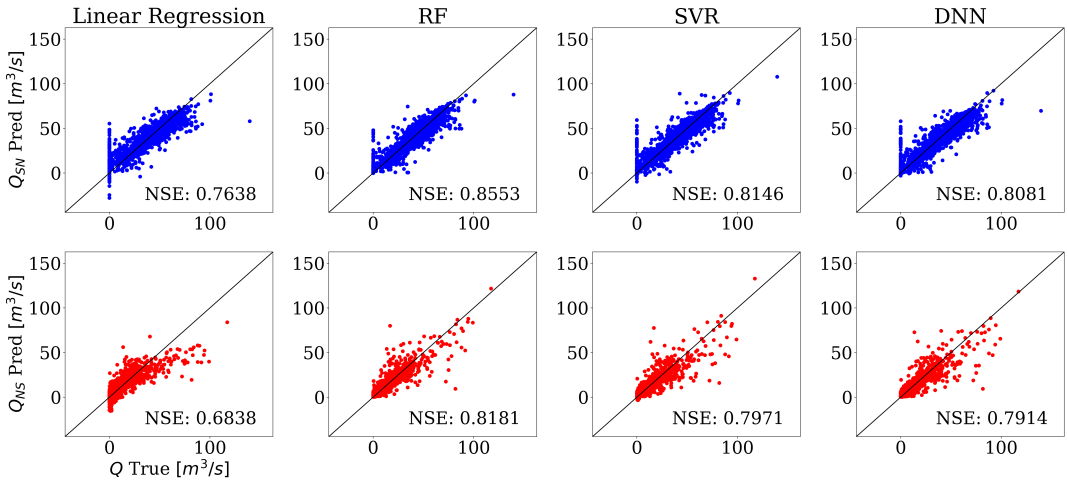

Fig. 3.6: Model performance compared on the test dataset from hydrological dataset using density.

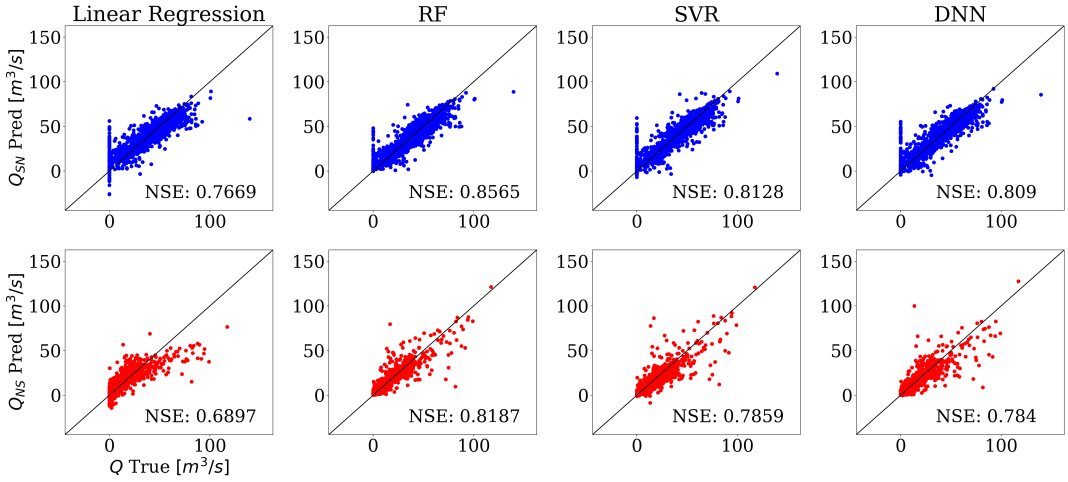

Fig. 3.7: Model performance compared on the test dataset from hydrological dataset using specific conductance.

Data-driven models depend on set conditions producing a set output, or something close to it. They learn relationships after seeing enough data that map an input to an output. Due to the fact that many different lake conditions can create a uni-directional it makes it difficult to create a good mapping for the uni-directional flow cases. From the

results shown in Figs. 3.6 and3.7, there is two main cases that can be seen where all models struggle to produce correct results. Where both of these issue cases are seen at the extremes where uni-directional flow dominates.

The first of these is in the SN flow prediction at the zero value. Here all models produce a large range of predicted values for the same known value. Here a physical knowledge of the GSL system is needed. For a uni-directional flow case to exist, there needs to be enough head pressure from the northern arm to overcome the southern arm pressure. This can be done by equalizing the WSE, since the rate of pressure increase in the northern arm is greater with its higher density. This case is not unique though and can be created by multiple lake conditions with a combination of wind shear influences and lowering lake elevation. Where these influences change the relative density and WSE difference between the two arms at the breach location. With these effects, uni-directional flow can exist for multiple configurations of inputs, and is not exclusive to a single case.

The same is true for the uni-directional flow case with SN flow dominating. Here, where NS flow is zero there is multiple configurations that lead to the southern arm hydrostatic pressure dominating the flow. Much like the uni-directional NS flow case each of the models struggle to produce the zero value given the different inputs. Here unlike the SN flow, there is a smaller range of predicted values when given the zero condition. Within the breach structure the NS flow is usually much less than that of the SN flow. This may contribute to the data being more well suited around lower values and having increased variability as higher values are experienced.

Taking the evaluation of these models one step further each of the models is evaluated on monthly test data recorded by USGS. This dataset consists of the monthly measurements recorded by USGS during site visits on a monthly to bi-monthly basis. This dataset consists of 30 site visits with known lake conditions taken over the years of 2018-2022. This dataset contains instances of WSE outside the range of tested WSE combined with density and specific conductance data also outside the range used for training the data-driven models. In machine learning each model is constrained by the dataset it is trained on. The larger

the dataset, and larger the range of data, the more robust a model can become. This USGS testing set is used to evaluate the performance of these models on real world data that can be outside the range given in the training datasets. The model results are shown in Table 3.9 with visual performance shown in 3.8 and 3.9.

Table 3.9: Statistical evaluation of data-driven models using USGS test dataset.

Model Type	Flow Dir.	Density/SpCond	NSE	RMSE	CC	PBIAS
LR	SN	Density	-0.9742	1.4051	0.8534	1045.6728
	NS	Density	-7.7395	2.9563	-0.0001	363.3940
	SN	SpCond	-1.2337	1.4946	0.8554	1121.2613
	NS	SpCond	-7.4769	2.9115	0.1997	385.9927
RF	SN	Density	0.1617	0.9156	0.8109	562.8407
	NS	Density	-37.0091	6.1652	-0.5087	705.8495
	SN	SpCond	0.0981	0.9497	0.7770	566.0585
	NS	SpCond	-38.1942	6.2605	-0.5385	717.1367
SVR	SN	Density	-2.8884	1.9719	0.3640	-657.8905
	NS	Density	-65.8715	8.1775	-0.3755	779.9938
	SN	SpCond	-1.2167	1.4888	-0.1286	338.4358
	NS	SpCond	-7.4500	2.9069	-0.1477	362.3022
DNN	SN	Density	-4.1230	2.2634	0.7747	1388.5421
	NS	Density	-258.8707	16.1205	-0.5897	1556.0037
	SN	SpCond	-0.0160	1.0080	0.7450	575.6174
	NS	SpCond	-266.3037	16.3494	-0.6272	1629.6986

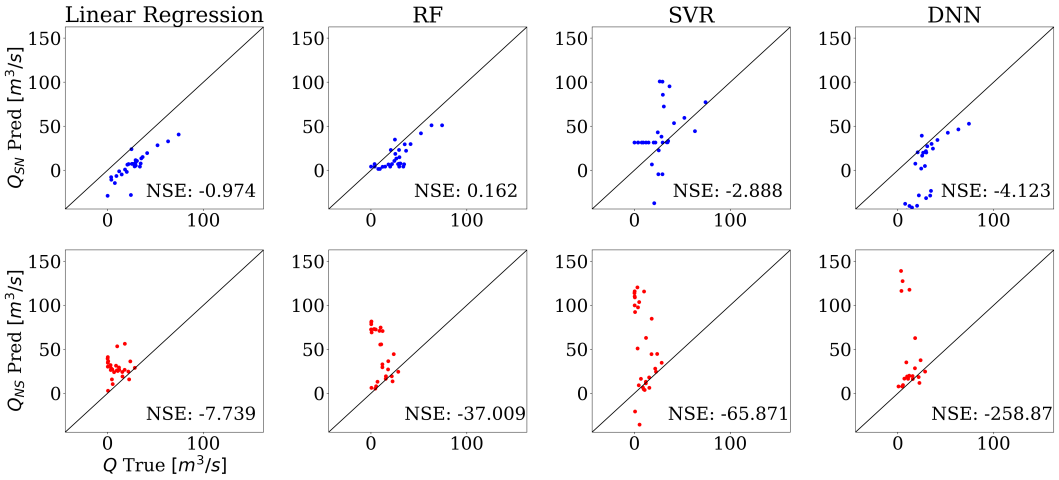

Fig. 3.8: Model performance compared on the test dataset from USGS test dataset using density.

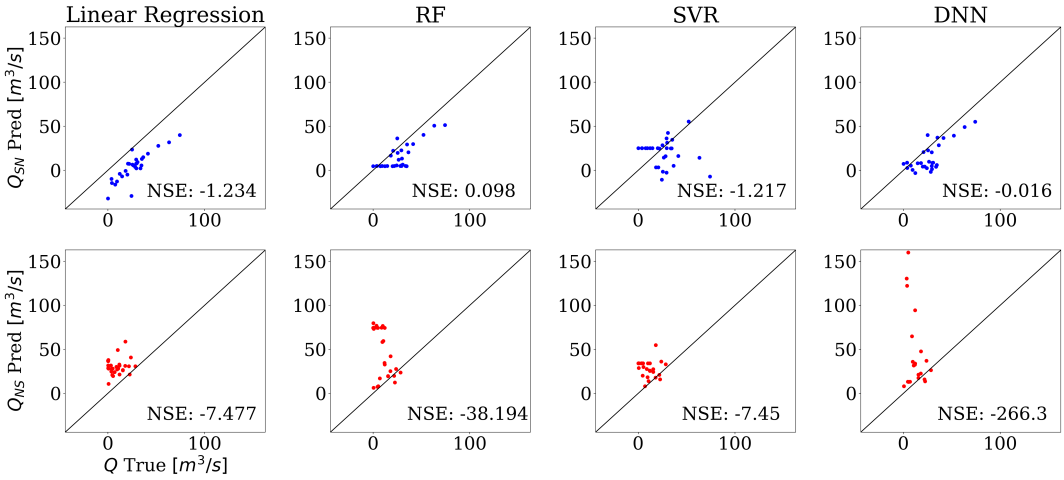

Fig. 3.9: Model performance compared on the test dataset from USGS test dataset using specific conductance.

From these results, each of the models breaks down as it sees values that are outside the range it was trained for. The most robust model being RF with an NSE value of 0.16, a value much lower than the performance seen during training. The most notable method visually is the LR method where the prediction follows the correct trend, though needs an

increase or decrease in the bias to create a more robust prediction. This result speaks to the robustness of LR and its application in hydrological applications. Where during evaluation on the hydrological dataset this method performs worse than all others, while preserving the expected trend during evaluation on this USGS dataset.

Given the results of this USGS test evaluation using monthly data, it is clear to see additional refinement is necessary to improve the robustness of any given method on flow prediction in the GSL. One of the most fundamental methods would be finding a different way to scale the data to preserve the variables distribution compared to nominal values of a training dataset. Standard Scalar is used in this study to try and preserve the relationship to nominal conditions. Using the relationship of the mean and standard deviation from nominal conditions is expected to improve the robustness of data-driven methods for out of dataset prediction. The opposite is seen in the USGS tested dataset, where originally each model performed well, and then performed worse for this test set. With all of this being said, the reduction in performance is expected because of the data-driven nature of ML models.

Further evaluation of data-driven is done by comparing these models to the current 1D model prediction used by Utah's Department of Natural resources (Utah DNR). This model is the adaptation of Holley's model from 1976 to the new trapezoidal NB channel. Since LR and RF show the most robust performance for flow prediction these models are compared. With the visual comparison shown in Fig. 3.10. The performance of the 1D model on the test dataset used for all other models is shown in Table 3.10. It should be noted that for flow prediction, the 1D model requires WSE for the north and south arms, either respective densities and the bottom height of the channel.

Table 3.10: Performance of Utah DNR's 1D model on hydrological test dataset.

Model Type	Flow Dir.	Density/SpCond	NSE	RMSE	CC	PBIAS
DNR 1D	SN	Density	-241.0154	15.5568	0.0571	-83.1029
	NS	Density	0.1610	0.9159	0.5868	37.0227

It is clear to see the current 1D model does not perform well for the new trapezoidal cross-section. The limitations of the 1D model are apparent when seen in the visualization compared to LR and RF shown in Fig. 3.10.

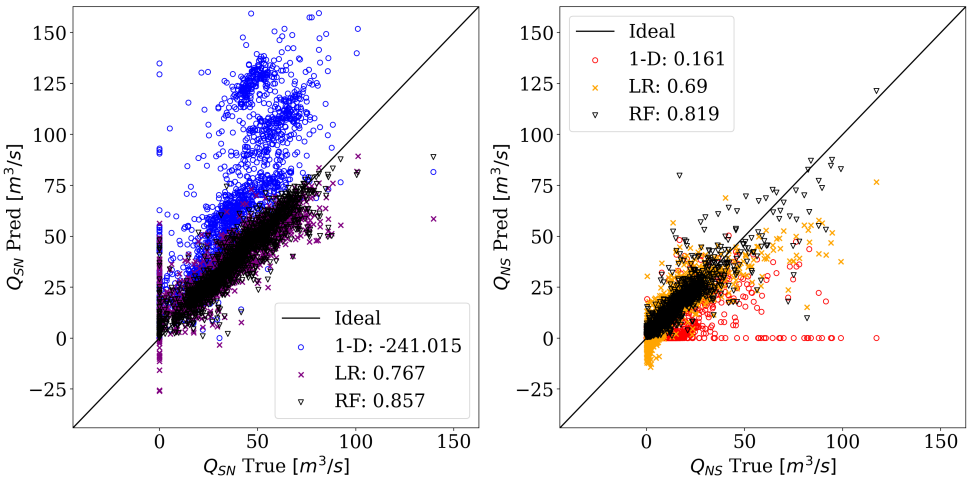

Fig. 3.10: Linear regression and random forest predictions compared to Utah DNR 1D model for SN and NS discharges.

The major issues seen in the 1D model are shown at higher discharge measurements. At lower SN and NS predictions the model is capable of producing a biased value that does follow the desired trend. Much like the data-driven models, the SN prediction is unable to produce correct results for zero discharge SN. Looking at the NS flow it is clear to see the model is over confident there is no NS flow going through the breach; with zero predicted discharge for true discharge flows up to approx. 120 cm/s. For conditions more generally seen, i.e. the mid-range flows, the model does not perform as far off from prediction as the higher end flows. Due to this fact, and the limitations of the 1D model it is clear to see the use of the current 1D model should be confined to the mid-range discharges for prediction.

3.2.4 Effects of input reduction in hydrological dataset

The comparison between the 1D model and the other data-driven models is not entirely a fair evaluation between models; given the increased input complexity of the data-driven models. To produce a more equivalent comparison between the LR, RF, and 1D models the input complexity is reduced in the LR and RF models to more consistently represent the data available to the 1D model. To do this first the wind data is removed and each model is reevaluated. The breach WSE is then removed to provide as equivalent input to all models as possible to the data available to the 1D model. For each of these models only the density is used to be more representative of what the 1D model uses for density information. The reduced model performance compared to the 1D model is shown in Table 3.11, with visual comparison of LR and RF shown in Fig. 3.11.

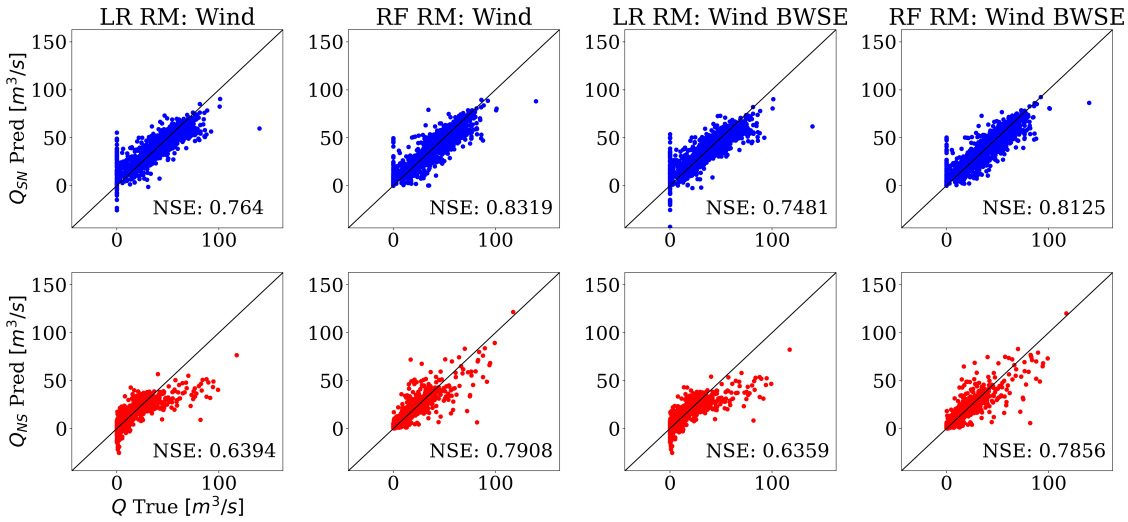

Fig. 3.11: Performance of LR and RF with reducing dimensionality of input.

Table 3.11: Statistical values of data-driven models reducing available inputs

Model Type	Flow Dir.	NSE	RMSE	CC	PBIAS
DNR 1D	SN	-241.0154	15.5568	0.0571	-83.1029
	NS	0.1610	0.9159	0.5868	37.0227
LR	SN	0.7638	0.4860	0.8741	2.4018
	NS	0.6838	0.5623	0.8277	22.8270
RF	SN	0.8553	0.3804	0.9250	1.4067
	NS	0.8181	0.4265	0.9047	4.8086
LR	SN	0.7640	0.4858	0.8743	2.5849
RM: Wind	NS	0.6394	0.6005	0.8008	29.4313
RF	SN	0.8313	0.4107	0.9118	0.5802
RM: Wind	NS	0.7914	0.4568	0.8899	2.9674
LR	SN	0.7481	0.5019	0.8656	6.0021
RM: Wind & BWSE	NS	0.6359	0.6034	0.7990	36.9333
RF	SN	0.8118	0.4338	0.9011	0.6665
RM: Wind & BWSE	NS	0.7836	0.4652	0.8858	4.3782

As input of each model is reduced performance does decrease in performance; though they still out perform the 1D model's performance. From this it is clear to see the data-driven models are capable of making well conditioned prediction even with reduced input comparable to the 1D input. The results shown in Fig. 3.11 show the problem at zero discharge persists for both models as input is reduced. When comparing the different performances each of the RF models exhibits a slight fanning out of the zero prediction issue with reduced input, but the rest of the prediction remains similar.

3.2.5 Bi-directional instance case study

Knowing that each of the models struggles to predict the uni-directional flow cases, how do the models compare if one only uses the bi-directional cases? Such a comparison allows one to see how the models are expected to perform for common operating conditions at the GSL NB. This bi-directional subdivision of the dataset is created by removing instances outside the 2σ bounds determined before, and then removing any remaining zero flow cases from the SN and NS flow cases. These reductions place emphasis on instances where bi-directional flow exists at the GSL. The developed models are tested on the new altered test dataset and the resulting performance is shown in Fig. 3.12 with full statistical results shown in Table 3.12.

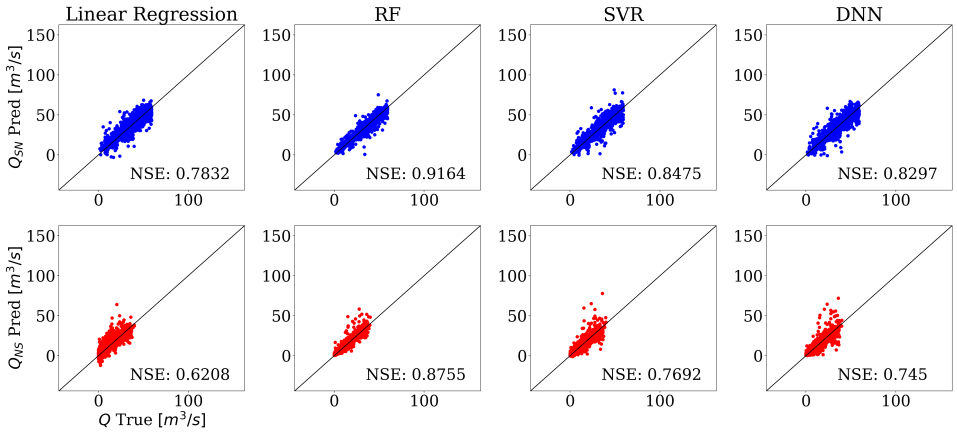

Fig. 3.12: Model performance compared on the test dataset from bi-directional test dataset using density.

Each of the models shows the desired trend with variable levels of banding across the ideal line. Here RF is the clear best performing method to predict the bi-directional flows as shown in Table 3.12. Compared to the other models the RF method has a lower band of prediction around the ideal. This behavior may be due to how RF makes its predictions, using the dataset to produce predictions given an ensemble of decision tree predictions. Since the data is used to predict the final value, the trend is consistent with the trend seen in the data.

Table 3.12: Data-driven models performance using hydrological bi-directional dataset.

Model Type	Flow Dir.	Density/SpCond	NSE	RMSE	CC	PBIAS
LR	SN	Density	0.7832	0.4656	0.8945	1.6038
	NS	Density	0.6208	0.6158	0.8325	75.3972
	SN	SpCond	0.7820	0.4669	0.8931	1.6024
	NS	SpCond	0.6269	0.6109	0.8401	69.9187
RF	SN	Density	0.9164	0.2891	0.9580	1.1901
	NS	Density	0.8755	0.3529	0.9411	18.3916
	SN	SpCond	0.9223	0.2787	0.9608	0.8547
	NS	SpCond	0.8772	0.3505	0.9419	18.2915
SVR	SN	Density	0.8475	0.3905	0.9252	-0.3124
	NS	Density	0.7692	0.4804	0.8909	8.3991
	SN	SpCond	0.8429	0.3964	0.9231	-0.2232
	NS	SpCond	0.7687	0.4809	0.8916	9.7615
DNN	SN	Density	0.8297	0.4126	0.9180	5.4073
	NS	Density	0.7450	0.5050	0.8838	16.9023
	SN	SpCond	0.8328	0.4089	0.9203	-3.0367
	NS	SpCond	0.7527	0.4973	0.8909	18.4921

Comparing the full data test dataset with the uni-directional flow to the test set using only bi-directional data is shown below in Table 3.13 with better performing values shown in bold. In this table most methods are have a mix of which dataset produces the better results.

Table 3.13: Data-driven model performance compared using bi-directional and total hydrological datasets.

			Full Data		Bi-Dir Data	
Model Type	Flow Dir.	Density/SpCond	NSE	CC	NSE	CC
LR	SN	Density	0.7638	0.8741	**0.7832**	**0.8945**
	NS	Density	**0.6838**	0.8277	0.6208	**0.8325**
	SN	SpCond	0.7669	0.8759	**0.7819**	**0.8931**
	NS	SpCond	**0.6897**	0.8313	0.6269	**0.8401**
RF	SN	Density	0.8553	0.9250	**0.9164**	**0.9580**
	NS	Density	0.8181	0.9047	**0.8755**	**0.9411**
	SN	SpCond	0.8565	0.9257	**0.9223**	**0.9608**
	NS	SpCond	0.8187	0.9051	**0.8772**	**0.9419**
SVR	SN	Density	0.8146	0.9026	**0.8475**	**0.9252**
	NS	Density	**0.7971**	**0.8932**	0.7692	0.8909
	SN	SpCond	0.8128	0.9016	**0.8429**	**0.9231**
	NS	SpCond	**0.7859**	0.8873	0.7687	**0.8916**
DNN	SN	Density	0.8081	0.8992	**0.8297**	**0.9179**
	NS	Density	**0.7914**	**0.8898**	0.7450	0.8838
	SN	SpCond	0.8090	0.8997	**0.8328**	**0.9203**
	NS	SpCond	**0.7840**	0.8860	0.7527	**0.8909**

This is true for all models except for the RF model. The RF model sees a substantial improvement in performance for the bi-directional dataset alone. This result shows that RF methods are expected to outperform other machine learning methods for the common bi-directional flow cases exhibited by the GSL.

3.2.6 Hybrid DNN

Knowing the 1D model is unable to perform well by itself makes it seem that it might be best to step away from this model. Instead, what if the outputs were instead coupled with a DNN method to act as a refinement layer on the prediction to the desired prediction. The combination of this data was integrated using two different methods, series and parallel HNN methods. Density is used as the density information for the network to be consistent with expected inputs to the 1D model. The similar architecture as the DNN is used with implementation of an changed input layer to either combine the 1D approximation with the current input (parallel) or to solely take the 1D outputs as the input to the HNN (series). The performance of the HNN is shown in Table 3.14 with visual results shown in Fig. 3.14.

Table 3.14: Performance of HNN models on hydrological test dataset compared to DNN baseline prediction.

Model Type	Flow Dir.	NSE	RMSE	CC	PBIAS
DNN	SN	0.8081	0.4381	0.8992	3.6342
Baseline	NS	0.7914	0.4567	0.8898	-15.3282
Parallel	SN	0.8066	0.4398	0.8986	3.3806
	NS	0.7930	0.4549	0.8908	-13.3937
Series	SN	0.6932	0.5539	0.8326	1.1895
	NS	0.6527	0.5893	0.8109	-36.6672

Increasing the complexity of the input vector using the 1D input produces comparative performance to the current DNN structure. There exists slight improvement to the percent bias in predicted value on the test set but little other improvement between the DNN and parallel HNN method. When comparing the series HNN there is a large decrease in performance. This result comes from the decrease in input dimensionality for the model to find representative methods to map input to output. The goal of the series model was to see if using the 1D model output could act as a type of transfer learning from a physically-based

numerical model to a data-driven model. Here it is clear to see this method decreases the expected performance and instead requires different implementation. When comparing the results visually one can see the physically-based numerical models influence in the system as shown in Fig. 3.14.

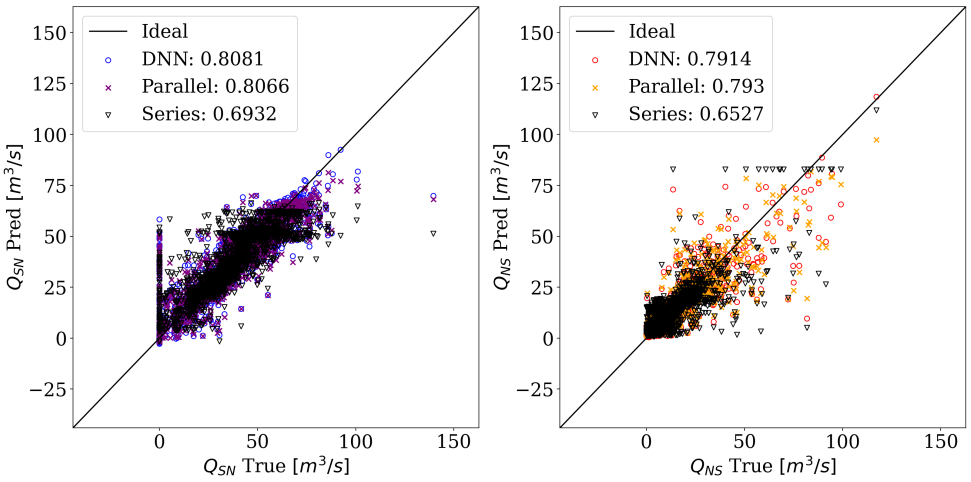

Fig. 3.13: Comparative performance of DNN, HNN parallel, and HNN series models for test data from development dataset.

Figure 3.14 shows the prediction trend of DNN dominates over the trends seen when using the 1D model alone (see Fig. 3.10). Here the model is capable of predicting the higher flow values. The spread of predicted values is the same amongst the different models both the DNN and HNNs. This trend unexpected due to the reduced complexity in the series HNN, though the different neural network structures seem robust enough to compensate for the reduced dimensional input.

Further investigation of the capabilities of HNN are shown using data collected during the 2021 year with predictions made using a velocity based DNN model. This is done to show the expected performance of the HNN for conditions that are held outside the range of training data. The results of this extrapolation study are shown in Fig. **??** with performance metrics shown in Table 3.15

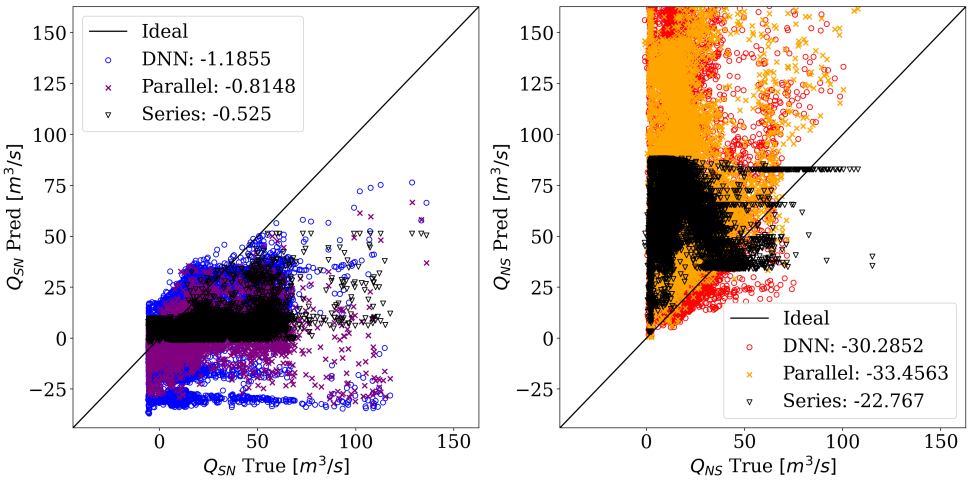

Fig. 3.14: Comparative performance of DNN, HNN parallel, and HNN series models for 2021 extrapolation dataset.

Table 3.15: Performance of HNN models on hydrological extrapolation dataset compared to DNN baseline prediction.

Model Type	Flow Dir.	NSE	RMSE	CC	PBIAS
DNN	SN	-1.1855	1.4783	0.4118	-185.1622
Baseline	NS	-30.2852	5.5933	0.2077	-1232.1102
Parallel	SN	-0.8148	1.3471	0.2804	-205.4909
	NS	-33.4563	5.8699	0.1149	-1881.2663
Series	SN	-0.5250	1.2349	0.4398	-184.8581
	NS	-22.7670	4.8751	-0.0569	-1671.1950

From the results of this extrapolation study show that HNN does provide slight improvement to the performance of the DNN. Although this is true all methods perform poorly and result in unreliable predictions. Due to this it can be seen that the application of a HNN does not work well for the GSL dataset to improve upon the ability of the DNN to generalize for data ranges outside of the training dataset.

Additional methods have been used by other researchers to implement physically-based numerical models by incorporating the solved equations into the loss function of the DNN. These type of networks are known as Physics Informed Neural Networks (PINN) [41] though they reside outside the scope of this study. The implementation of these PINNs seeks to use a weighted loss function to combine the regular loss term like MSE with the physical equations that govern the flow, and then updated the weights and biases using that weighted loss function. Such a method increases the computational cost of model development, where the goal here was to see if performance could be increased by using raw outputs of a physically-based numerical model.

3.2.7 Velocity Dataset Data-driven Models

The implementation of data-driven models for the hydrological dataset and their respective performances are shown in the previous sections of this study. With this data there is a key issue, the dataset is dependent on four different instruments at USGS site locations 10010024-10010027 all recording the proper data. Additionally, a monthly density or specific conductance measurement must be recorded for the time period in question. These data requirements limit the versatility of the produced model because if any of these measurements are not record a full input vector can not be generated.

Instead of relying on all of these data collection sites, one can instead use the velocity column data from USGS's ADCP located at site 10010025. This data instead requires a single measurement station to be operational to produce the required input vector to a data-driven method. Using this Velocity dataset the LR, RF, SVR, and DNN models are redeveloped and their respective performances are shown in Table 3.18, with model configurations listed below.

3.2.8 Model Configurations

Linear Regression

The resulting linear regression model using the velocity dataset is shown in Eq. 3.10 and 3.11 for SN and NS flow predictions respectively.

$$\begin{aligned} Q_{\mathrm{SN}} = {} & 0 \cdot \mathrm{Cell0} + 0.5585 \cdot \mathrm{Cell1} + 0.2144 \cdot \mathrm{Cell2} - 0.1982 \cdot \mathrm{Cell3} \\ & + 0.3907 \cdot \mathrm{Cell4} - 0.6506 \cdot \mathrm{Cell5} + 0.1436 \cdot \mathrm{Cell6} + 0.0687 \cdot \mathrm{Cell7} \\ & + 0.3971 \cdot \mathrm{Cell8} - 0.2929 \cdot \mathrm{Cell9} + 0.3381 \cdot \mathrm{Cell10} \end{aligned} \quad (3.10)$$

$$\begin{aligned} Q_{\mathrm{NS}} = {} & 0 \cdot \mathrm{Cell0} - 0.8253 \cdot \mathrm{Cell1} + 0.8174 \cdot \mathrm{Cell2} - 0.3868 \cdot \mathrm{Cell3} \\ & - 0.1471 \cdot \mathrm{Cell4} - 0.2135 \cdot \mathrm{Cell5} + 0.0960 \cdot \mathrm{Cell6} + 0.0366 \cdot \mathrm{Cell7} \\ & - 0.1061 \cdot \mathrm{Cell8} + 0.0426 \cdot \mathrm{Cell9} - 0.4817 \cdot \mathrm{Cell10} \end{aligned} \quad (3.11)$$

Here it is seen that most velocities are considered with comparative weight except for Cell 0. Since Cell 0 is the no slip boundary condition, it is understandable it is not used in the LR model because the value does not fluctuate. Therefore, this velocity is not considered in the model because it does not capture any of the variation seen in the dataset.

Random Forest

Using the velocity data, the RF model configuration uses a max depth of 20, 3 minimum samples per leaf node and 500 decision trees (see Table 3.16). The variable importance used in the RF model is shown in Fig. 3.15.

Table 3.16: Velocity based dataset RF model configuration.

Max Depth	Min Samples	N Trees
20	3	500

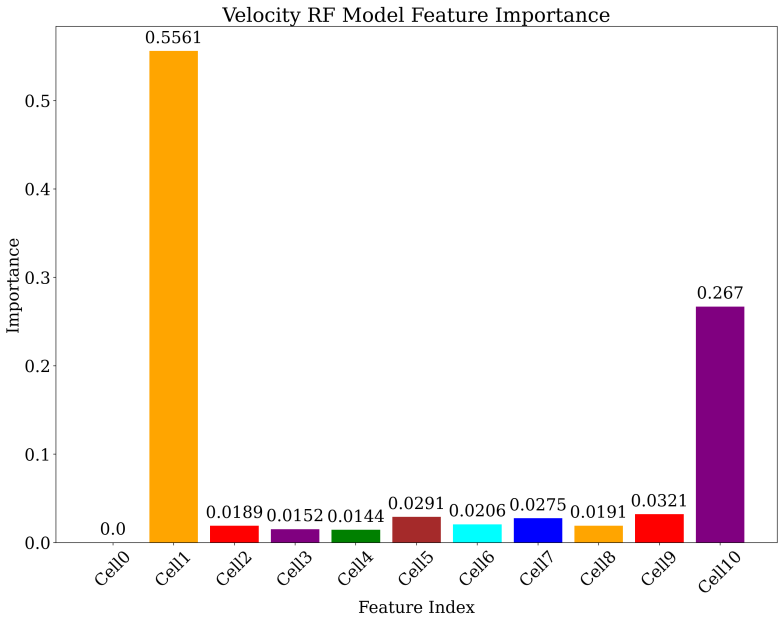

Fig. 3.15: Variable importance of the velocity based random forest model showing importance of bounding cells to prediction.

It is interesting to note from Fig. 3.15 the two major contributors to the data separation used in the RF method are Cell1 and Cell10; the ADCP's bottom and top cells. The only cell not considered is Cell0 because all samples share this common value of zero. Cells 2-9 are each considered with comparable importance, so from this analysis the main values to be considered when separating the values in the RF model are the boundary cells from the ADCP.

Support Vector Regression

With the velocity based SVR there is still a requirement to create two different models to produce the two flow predictions through the NB. From the grid search method a best performing model was developed. The model configurations for γ, C and kernel are shown in Table 3.17.

Table 3.17: Configurations for support vector regression using velocity dataset.

Flow Dir.	kernel	C	γ
SN	RBF	10	1
NS	RBF	10	0.1

As seen before, each of the models uses the RBF kernel with the larger Gaussian bandwidth required in the NS prediction.

Deep Neural Network

The DNN developed uses a [100,32,100,50,25] configuration with a 11 dimensional input and 2 dimensional output. This DNN structure gives 14,057 trainable parameters for the model, with the best performing activation function being ReLU.

3.2.9 Model Performance

Assessment of the overall performance of the developed models using the study performance parameters are shown in Table 3.18.

Table 3.18: Statistical values of data-driven models using velocity test dataset comparing model performance.

Model Type	Flow Dir.	NSE	RMSE	CC	PBIAS
LR	SN	0.6153	0.6202	0.7904	53.1379
	NS	0.8449	0.3939	0.9194	207.3436
RF	SN	0.8649	0.3676	0.9300	-1.8816
	NS	0.8907	0.3305	0.9438	-40.0168
SVR	SN	0.8497	0.3876	0.9219	1.3283
	NS	0.8918	0.3289	0.9444	-141.4931
DNN	SN	0.8642	0.3685	0.9297	-4.1568
	NS	0.8923	0.3282	0.9463	-497.5495

Comparing these results to those shown in Table 3.8 there is a clear improvement to data-driven model performance when using the velocity dataset. There are multiple contributing factors that can account for this increase in performance. One possible reason being the dimensionality increase from the velocity input data. Every data-driven model relies on creating relationships between input parameters, heavily reliant on the dimensionality of that input.

The difference between methods is how those input dimensions are utilized for decision making. Random forest develops a separation scheme using each dimension based on value. Linear regression uses each dimension to create a single equation to predict flow. Support vector regression uses the high dimension to create a hyperplane to explain the data and create an equation for prediction. Where methods like DNN use the increased dimensions to develop additional relationships between input and output metrics through their layer structure. Therefore, increasing the dimension can improve the overall prediction of the model, but this must be considered carefully. One can increase the dimension of input by increasing the dataset size without adding any additional information.

An example of this flaw is seen in the inclusion of Cell 0 for flow prediction using the velocity dataset. When physically modeling the system numerically, it makes sense to include the boundary condition to ensure the proper flow is produced, whereas using data-driven models it doesn't make sense because the value explains no variability. Without any variation in this value each model is unable to create a distinction between points using this data. Therefore, this data is ignored in the dataset and instead increases the computation cost without yielding any additional data or benefit.

While this is true, there are cases one can increase the dimensionality of input by creating derived parameters for data-driven model by combining known data inputs. This is analogous to non-dimensionalization in fluid mechanics; where using base data you can create more refined variables that are shared amongst multiple configurations. Using these methods to create derived data, it reduces the computational burden on data-driven models

to formulate these relationships itself. Instead they are given as data input at the beginning allowing the model to create other relationships not as readily apparent.

There are many ways to increase the dimensionality of a dataset, here the dimensionality is increased by using a different set of features available in the compiled dataset. With the increased dimensionality the models improve and the improvement is clearly seen in Table 3.18. Though this new dataset improves the overall performance of most data-driven models, there is a decreased performance in the LR method. Additionally, looking at the visual performance of the network, the same prediction issues persist for flow predictions at extreme flow events (see Fig. 3.16).

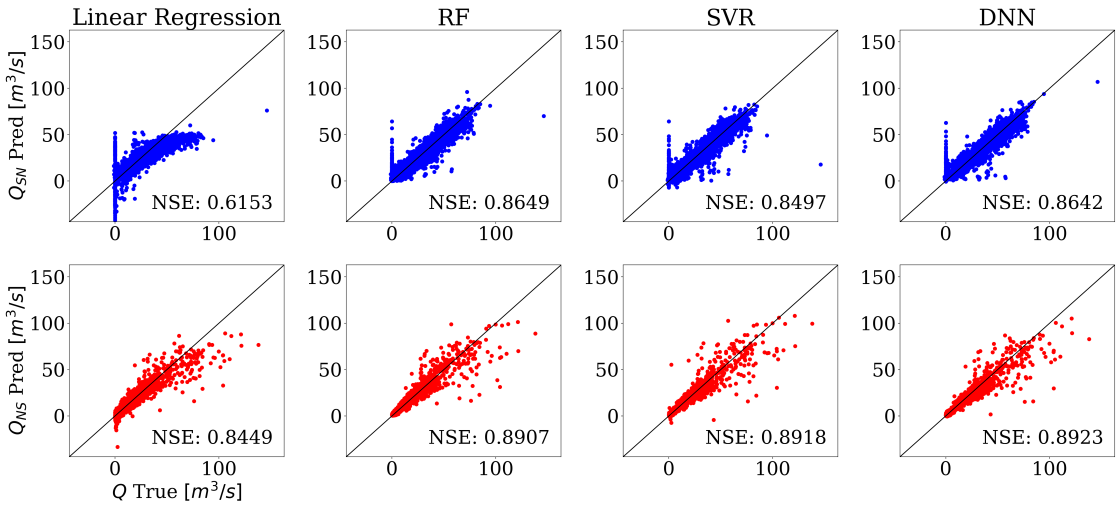

Fig. 3.16: Model performance compared on the test dataset from velocity dataset.

Here it is seen the ML models using the velocity dataset are still unable to accurately predict the SN discharge during uni-directional NS flow cases. Overall though there is refinement of the prediction along the ideal trend, for both discharge predictions the spread of values is reduced at higher discharge values. The improvements in performance show the velocity dataset is more well suited for flow prediction at the GSL NB using data-driven methods. These data-driven models are then compared to monthly measurements conducted by USGS and the results are shown below in Table 3.19 and Fig. 3.17

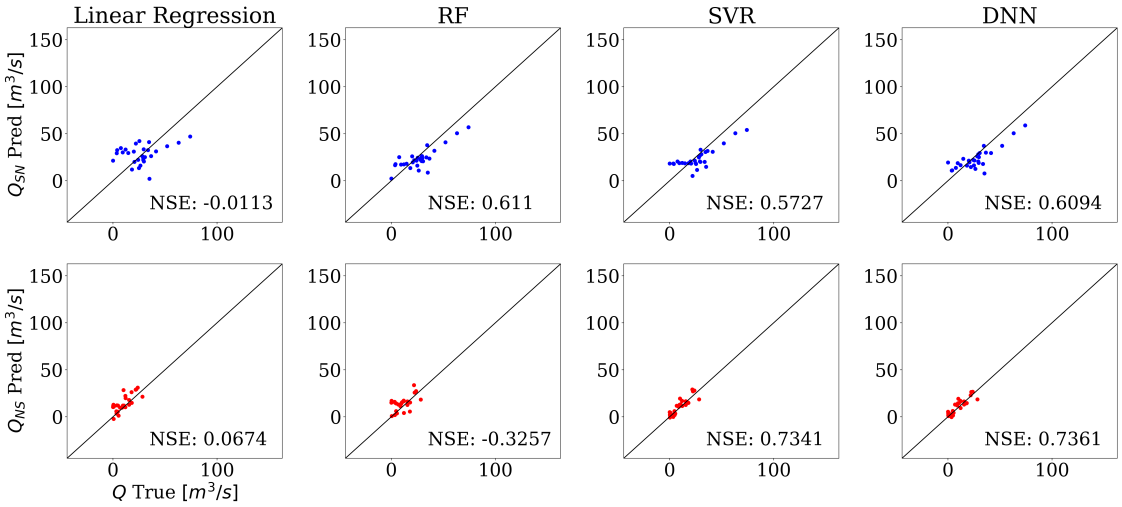

Fig. 3.17: Model performance of velocity based ML methods compared on the test dataset from USGS test dataset.

Table 3.19: Statistical values of data-driven models using USGS test dataset

Model Type	Flow Dir.	NSE	RMSE	CC	PBIAS
LR	SN	-0.0113	1.0056	0.2989	-68.5238
	NS	0.0674	0.9657	0.7393	92.3748
RF	SN	0.6110	0.6237	0.8163	157.5445
	NS	-0.3257	1.1514	0.4747	92.9523
SVR	SN	0.5727	0.6536	0.7955	146.7213
	NS	0.7341	0.5156	0.8909	25.2771
DNN	SN	0.6094	0.6250	0.8196	169.0855
	NS	0.7361	0.5138	0.8793	24.6457

These results show the velocity based data-driven models are more well suited to applications outside the dataset. There is a caveat here though, USGS does not record the velocity profile at the time of measurement using the ADCP because it is under maintenance during these measurements. Due to this, the measurement had to be taken from

the closest instance to the measurement time. This is done assuming field measurements can only be handled when lake conditions are calmer. Therefore, the lake is assumed to be in a quasi steady state. With the inferred data from the full dataset the relationships are similar to those seen during development. This knowledge can be why the predicted values are more well conditioned during this test. The velocity based models are not compared to the 1D prediction because the required data for the 1D model is not contained in the velocity dataset, and the 1D model's performance was already assessed in the hydrological dataset of this study.

3.2.10 Limitations of Data-driven Models and Monitoring Site Data

With any data-driven models there are limitations to how they perform, and what they can be used on. Here it was seen that both datasets performed worse than training performance when used on the USGS monthly test dataset. This difference in performance is due to the relationships contained in the data. The USGS monthly test data lies outside the training dataset values for the hydrological dataset, with the velocity dataset models performing closer to expected values due to the velocity data available in the USGS test dataset being from the development set of the velocity models. With these results it is clear to see training performance is expected only for data that lies within the data similar to trends used for training the model. Therefore, the larger and wider datasets is, the more robust a data-driven model is expected to be for that prediction.

One of the major limitations one needs to be aware of in the GSL monitoring data is the variability within sampled values in each of the sampling locations. These measurements are made by field instruments that have inherent variability in their measurement. This means for any given measurement there is a band of values the instrument can output as the desired measurement due to the uncertainty from the instrument itself.

Due to this uncertainty in field measurements, you will see variability in the overall prediction from any of the ML methods; because fluctuations in the measurements will result in fluctuations in the prediction. Therefore, one must consider how the models compare on a performance level, and also assess each one's sensitivity to changes in input. Doing

so one can assess which method is the most robust to real world monitoring with noise and uncertainty in measurements collected. Where all methods might give comparable performance metrics for a given data set, but have varying levels of sensitivity to any one input's value range.

Another major limitations of development of data-driven models for GSL flow prediction is when changes are made to the physical structure affecting the available dataset and therefore model performance. The major control structure used in the GSL NB channel is the north side control berm. This berm contributes to the amount of NS flow allowed into the southern arm. The berm helps to maintain the salinity of the southern arm of the lake. All data-driven models developed in this study are formulated for an as build berm condition, known as the 0ft berm at a height of 4183 ft above NVGD. As the salinity of the southern arm rose in recent years, the berm was raised to 4ft in February 2023, and later raised to 9ft in July 2023 to restrict the NS flow. These changes influences the governing physics by physically altering the channel geometry.

Due to the change to the actual physical system, the data sets developed for a 0ft berm can no longer be used for current GSL predictions because the systems are no longer consistent. When the physical system is changed all data previously collected can not be used, because the flow dynamics are now changed. Using the current datasets to predict flow in a system that is no longer the same will result in erroneous values. If further data-driven models are expected to be developed a new data set must be generated with correct physical features considered. In order to effectively train data-driven models there needs to be a large enough dataset to train the model on what truth is. The USGS flow prediction data is only recorded until January of 2020, when the 0ft berm existed. USGS monthly measurements do exist after the berm was raised, but there are only 13 available instances. With a limited number of data points, data-driven models do not make sense due to lack of information to properly train the models. Instead there needs to be an additional method to predict the flow through the breach unrestricted by a dataset focused on the physical methods of flow through the NB.

CHAPTER 4

OBJECTIVE 4: PHYSICALLY-BASED NUMERICAL MODELING OF GSL NB FLOWS USING INDEX AND SHALLOW WATER BASED MODELS

4.1 Methodology

The following chapter will describe the application of physically-based numerical models for rapid prediction of GSL NB flows. This chapter will explain the methodology for developing the physically-based numerical models, how these models are applied to the GSL problem, and their accompanying results. The physically-based numerical models of this study are the Generalized Area-Based Index Model (GABI) and the Steady Shallow Water Exchange-flow Model (SSWEM); a new derivation of the 1D equations of Holley and Waddell [1] for a trapezoidal breach structure.

4.1.1 Generalized Area Based Index Model

At a fundamental level, the discharge of a system can be determined as long as the velocity and geometric characteristics of the flow system are known. Using this data, one can simply multiply the velocity by a given cross-section to estimate a discharge value. The Generalized Area-Based Index Model (GABI) predicts SN and NS discharges based on the NB geometry and the velocity measurements available via USGS data site 10010025; where the geometry is provided by bathymetry measurements of the NB. For USGS velocity measurements the water column is divided into 10 separate cells where the average velocity is recorded by the ADCP for each of the cell sections (see Fig. 4.1). A zeroth cell is added to cover the blanking distance between the ADCP and the first available velocity cell.

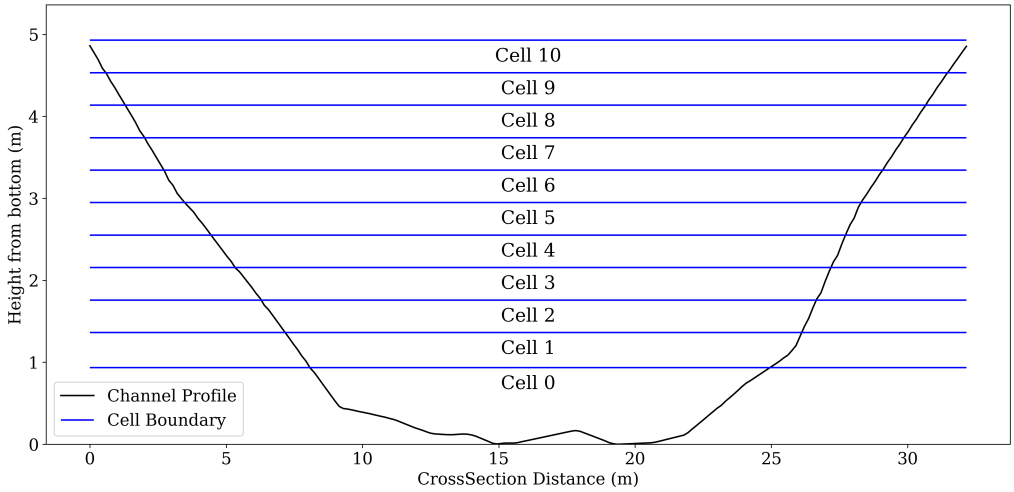

Fig. 4.1: New Breach geometry and velocity cell locations utilized by GABI for area formulation.

With the geometric data available, the GSL system still presents an interesting issue because of the two-layer flow characteristics give the three distinct flow cases; bi-directional flow and uni-directional flow either SN or NS. Therefore, in order for GABI to estimate the flow an interface location (H_I) between the two flow layers must first be estimated. GABI predicts the interface location using a simplified hydrostatic analysis to determine the point at which the north and south hydrostatic pressures are equal. The hydrostatic interface location is estimated using Eqs. 4.1, 4.2 and can be seen in Fig. 4.2.

$$H_I = 0.5 * (H_N + H_S) - \left(\frac{\rho_S}{\rho_N - \rho_S}\right) * \Delta WSE \tag{4.1}$$

$$\Delta WSE = H_S - H_N \tag{4.2}$$

Where H_I is the height of the interface above the channel bottom, H_S and H_N are the water heights of each arm above the channel base, and ρ is the respective density of each side of the lake.

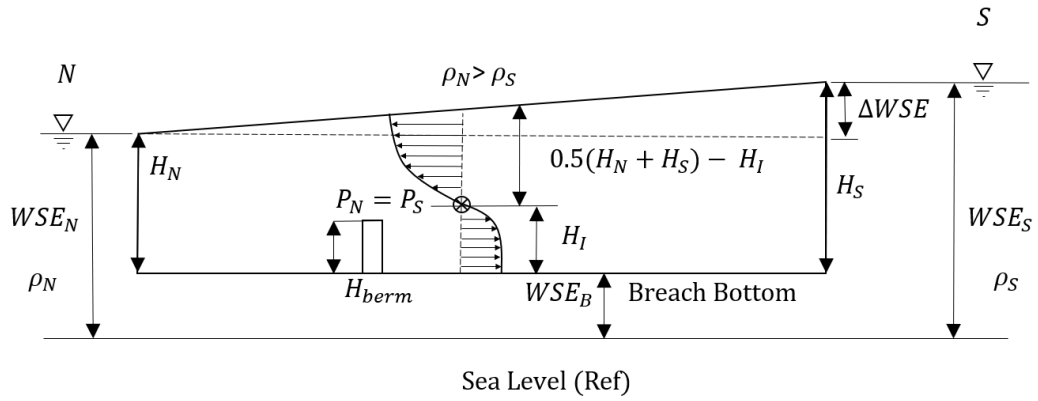

Fig. 4.2: Great Salt Lake New Breach cross-section for hydrostatic assessment, where flow above the pressure point is SN directional, and anything below it is assumed to be NS flow.

Using the height of equal pressure (H_I) from the bottom, anything above this point is assumed to be SN flow, and anything below this point is NS flow. Given that the interface location will not fall exactly on a dividing line between cells, the GSL cross-section is further subdivided into N number of sub sections to increase resolution of the geometry to divide the two flow layers. This refined geometry presents a problem for the velocity measurements. The velocities measures taken by the ADCP are a cell average and do not give exact locations of measurement for each cell. Therefore, in this study each velocity from the ADCP is set to the mid line of the respective cell and linear interpolation is used between points to assess the appropriate velocity for each of the subdivided cross-sections.

It is important to note that changes in the berm geometry have a major influence on the flow. The berm structure is assumed to be a non-permeable obstruction to the flow. Due to this, heights below the berm structure are not considered to contribute to the overall discharge. Therefore, for GABI, only cells above a set base height are considered. For this study the base height is set to the berm elevation. From analysis of different simulations and monthly measurements it is seen that uni-directional flow exists most commonly when the pressure point resides 0.3048 m (1 ft) below the base elevation. Therefore, the discharge calculation for GABI is discretized from this height location to the maximum water surface elevation using N subsections. If the pressure point is below this location, uni-directional

SN flow is assumed, and if the pressure point is at or above the maximum water surface elevation a uni-directional NS flow is assumed. Bi-directional flow exists when the pressure point location H_I satisfies $H_I \geq H_{berm} - 0.3048m$. Lastly a weir style discharge occurs when only one water elevation is above the berm elevation.

GABI uses five different classifications for discharge calculation with conditions listed in Table 4.1. The five flow cases are; no flow, bi-directional flow, SN flow, NS flow, and weir discharge flow. The no flow condition exists when neither north or south WSE are above the base elevation set for the model.

Table 4.1: Flow case scenarios

Flow Cases	H_I	H_N	H_S
No Discharge	NA	$< H_{berm}$	$< H_{berm}$
Bi-directional	$\geq H_{berm} - 0.3048m$	$> H_{berm}$	$> H_{berm}$
SN Flow	$\leq H_{berm} - 0.3048m$	$> H_{berm}$	$> H_{berm}$
NS Flow	$\leq H_{berm} > max(H_N, H_S)$	$> H_{berm}$	$> H_{berm}$
Weir Discharge	NA	$< H_{berm}$	$> H_{berm}$

Note: Weir discharge must have only one lake side elevation is greater than H_{berm}

The weir discharge case is calculated using Eq. 4.3. For the specific case of the GSL, the C_D value is determined using the weir geometry using the geological survey circular 397 [42]. Given the berm trapezoidal shape and the respective $\frac{h}{L}$ value given the berm width and height being a maximum of 0.133 the CD value for this application is 2.9. The length is approximated using the berm height of 1,277.72 m (4192ft) given the February 2023 change in berm elevation, giving an estimate of 30.48m (100ft). For the discharge calculation all measurements must be in English units for use in 4.3. The GSL weir discharge calculation is shown in Eq. 4.4

$$Q = C_D * L * H^{3/2} \tag{4.3}$$

$$Q = 2.9 * 100 * H^{3/2} \tag{4.4}$$

Where H here is the height of the water upstream relative to the top of the control berm measured in ft (see Fig. 4.3).

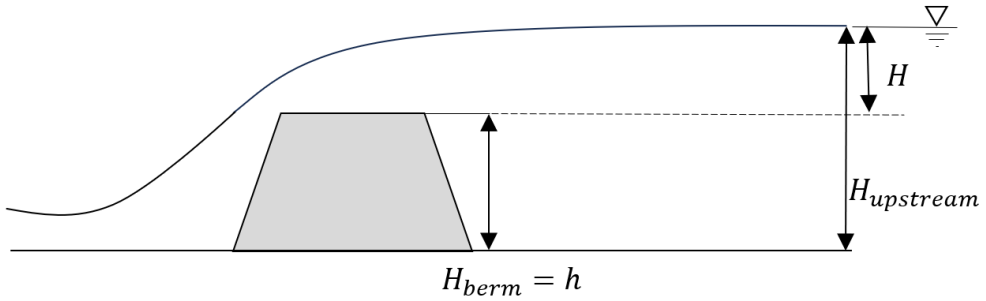

Fig. 4.3: Longitudinal view of weir discharge structure for obtaining required H of discharge calculation dependent on the associated berm height.

Therefore, knowing the geometry, interface location, base height, and having a set of velocities for each cell, the GABI model can predict the expected discharge. The solving procedure for flow predictions using GABI is as follows:

1. Obtain system information for: velocity, geometry, base height, lake elevations, and desired number of subdivisions
2. Discretize domain from base height to maximum surface elevation by N subdivisions
3. Solve for the H_I location using Eq. 4.1
4. Assess necessary flow case
5. Iterate through all subsections to calculate the discharge
6. Classify discharge to appropriate flow direction based on cell height relative to H_I
7. Repeat process for each time instance of input dataset

4.1.2 Steady Shallow Water Exchange-flow Model

The exchange flows of the GSL are not only driven by a difference in water surface elevation, but have dependencies on the local density gradient between the two arms of the lake. The higher density in the northern arm of the Lake allows for a bi-directional flow to exist driven by an increase in hydrostatic pressure overcoming pressure from the south arm within the NB channel. To model this behavior the Steady Shallow Water Exchange-flow Model (SSWEM) is developed to consider these flow dynamics.

To understand the dynamic fluid transport, this study needs to be able to understand the governing equation of the flow. Similar to other hydraulic systems, one can resolve the system by using momentum and energy transport. Before utilizing the momentum and energy transport of the system, first the geometry of the NB channel is identified to understand the method by which the lake transports these flows.

The NB is an open channel with a trapezoidal cross-section (see Fig.4.4). In the original flow model for the GSL created by Holley and Waddell [1] cross-section is a rectangular cross-section of a concrete box culvert (see Fig.4.5). With the opening of the NB, these original equations continue to be used, with adaptations to compensate for the changed cross-section of the channel while maintaining the same flow assumptions. The main flow assumptions used by Holley and Waddell were.

1. The culvert is a simple box culvert
2. The culvert never flows full
3. The flow is steady
4. The flow in each layer is one-dimensional and gradually varied
5. There is no entertainment across the interface

These assumptions allow the system to resolve the expected discharge for given Lake conditions of north WSE, south WSE, and their respective densities. This flow model uses an iterative scheme, assuming an initial discharge for a given layer, solving for the boundary

conditions, then checking the accuracy of those boundary conditions to the known lake conditions. If this set of lake conditions is incorrect, the conditions are then adapted and run again until the system of equations converges on a desired flow regime with its associated discharge measurement. While this method works well for box culverts, the adaptation of the original derivation does not hold for the new channel geometry. Due to this fact, the governing equations are derived for the new trapezoidal cross-section.

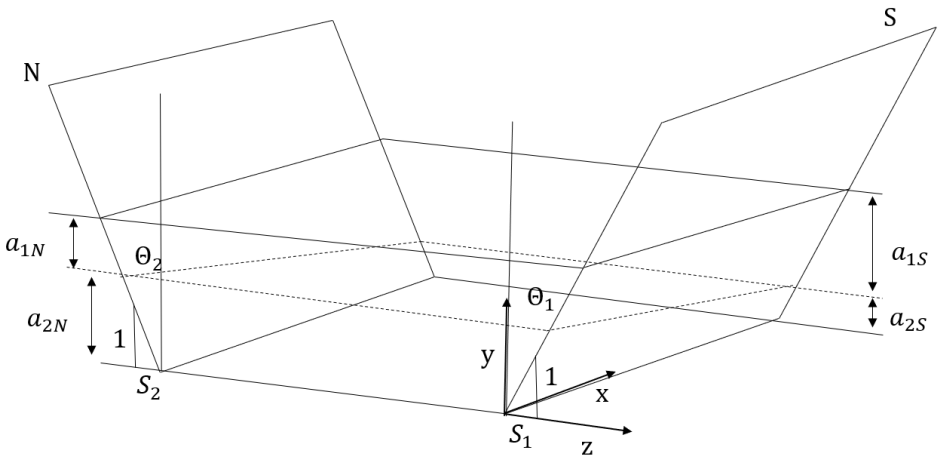

Fig. 4.4: Great Salt Lake, New Breach simplified trapezoidal cross-section.

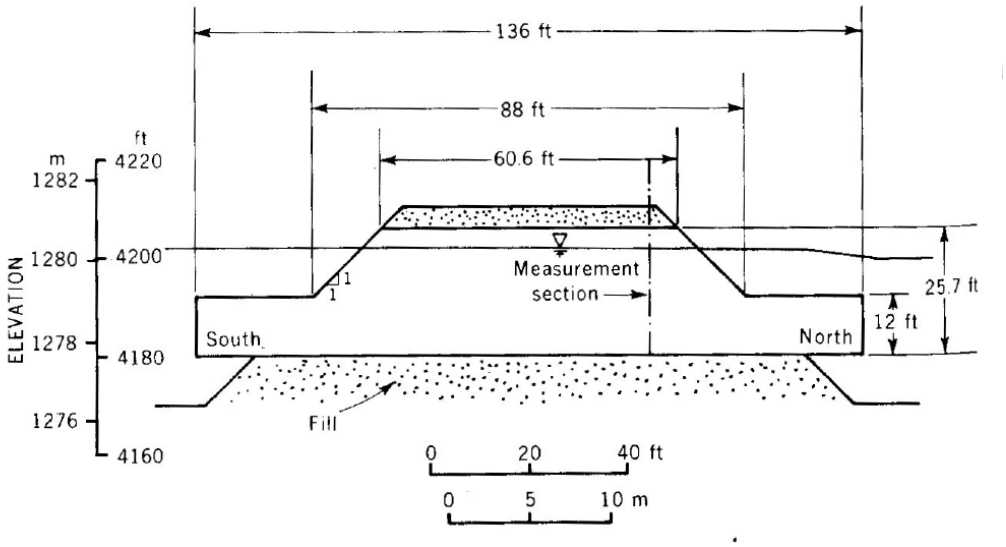

Fig. 4.5: Longitudinal view of box culvert, from the work of Holley and Waddell [1].

Knowing the original derivation does not generate true governing equations for the flow in a trapezoidal cross-section, this study can begin the development of SSWEM using the conservation of linear momentum. SSWEM uses the following assumptions to resolve the system.

1. The flow is incompressible
2. The channel never flows full
3. The flow is steady
4. The flow in each layer is one-dimensional and gradually varied
5. There is no entrainment across the interface
6. The channel floor is level

Using the conservation of linear momentum yields Eq. 4.5 for the base of the derivation.

$$\rho_i \left(\frac{du_i}{dt} + u_i \frac{du_i}{dx} \right) = -\frac{dP_i}{dx} + \frac{d\tau_{xy}}{dy} + \frac{d\tau_{xz}}{dz} + \rho_i g S_{0_i} + \rho_i g E_{l_i} \tag{4.5}$$

Where i denotes the layer of the flow, 1 = upper layer, while 2 = lower layer, and u is the velocity of the layer at any given point, τ is the shear stress in a given plane. S_0 is the slope of the channel, for this application it is assumed zero but is included for completeness of Eq. 4.5. Lastly, E_l is the equivalent distributed entrance loss for a given length of the channel. Simplifying given the assumption of a level channel, i.e. $S_0 = 0$, yields Eq. 4.6 where the slope S_0 no longer needs to be considered. Though it should be noted that in the upper layer formulation the change in lower layer height must be considered as this slope parameter.

$$\rho_i \left(u_i \frac{du_i}{dx} \right) = -\frac{dP_i}{dx} + \frac{d\tau_{xyi}}{dy} + \frac{d\tau_{xzi}}{dz} + \rho_i g E_l i \tag{4.6}$$

The entrance loss is not considered as a single loss for SSWEM, nor the original derivation due to a smooth water surface in the channel. When looking at the GSL NB channel, the flow has a smooth transition from one side to the other, without large hydraulic jumps in the water surface at the entrance or the exit of the channel. Due to this fact, the entrance losses is distributed over the length of the channel based on the discretization of the domain. The entrance loss formulation is shown in Eq. 4.7, with the distributed entrance loss is shown in Eq. 4.8

$$H_{Ei} = K_i \frac{U_i^2}{g} \tag{4.7}$$

$$E_{li} = \frac{H_{Ei}}{L} \tag{4.8}$$

Where H_E is the entrance loss, K is the loss factor dependent on the inlet structure, and L is the length of the NB channel in the X direction. Using the assumptions for this study and the simplified linear momentum equation (Eq. 4.6), the upper and lower layer governing equations are solved. Integrating this equation over the y and z domains of the sectional slice yields Eq. 4.9, with the known cross-section shown in Fig. 4.6. Where the flow area calculation is shown in Eq. 4.10, given the assumption that both side slopes are equal resulting in $\theta_1 = \theta_2$, allowing for the slope to be simplified in Eq. 4.11.

$$A_i \left(\rho_i \left(u_i \frac{du_i}{dx} \right) + \frac{dP_i}{dx} \right) = \frac{2a_i}{cos(\theta)} \tau_{wi} + B_{0i} \tau_{bi} + (B_{0i} + Sa_i) \tau_{si} + \rho_i g A_i E_{li} \tag{4.9}$$

$$A_i = a_i \left(B_{0i} + \frac{S}{2} a_i \right) \tag{4.10}$$

$$S = S_1 + S_2 \tag{4.11}$$

$$\Theta = tan^{-1}(S/2) \tag{4.12}$$

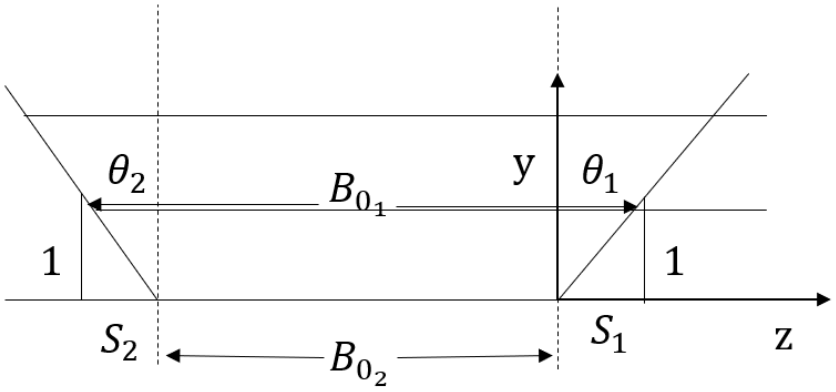

Fig. 4.6: cross-sectional view of GSL NB channel in the YZ plane.

Where A is the flow area, a is the given layer height, and τ_b and τ_s are the lower and upper shear stresses of the cross-section, τ_w being the wall shear stress, and P_i is the wetted perimeter of the flow. Using these equations, the upper and lower layer linear momentum is resolved as shown in Eqs. 4.13, 4.14.

$$\rho_1 g A_1 \left(\frac{u_1}{g} \frac{du_1}{dx} + \frac{da_1}{dx} + \frac{da_2}{dx} \right) = \frac{2a_1}{cos(\theta)} \tau_{w1} + B_{01} \tau_{b1} + \rho_1 g A_1 E_{l1} \tag{4.13}$$

$$\rho_2 g A_2 \left(\frac{u_2}{g} \frac{du_2}{dx} + (1-\epsilon) \frac{da_1}{dx} + \frac{da_2}{dx} \right) = \frac{2a_2}{cos(\theta)} \tau_{w2} + B_{02} \tau_{b2} + (B_{02} + Sa_2) \tau_i + \rho_2 g A_2 E_{l2} \tag{4.14}$$

$$\epsilon = \frac{\rho_2 - \rho_1}{\rho_2} = \frac{\Delta \rho}{\rho_2} \tag{4.15}$$

Dividing these equations by their respective cross-section, gravity, and layer density yields the equations for the velocity and layer height change over the x domain.

$$\frac{u_1}{g}\frac{du_1}{dx} + \frac{da_1}{dx} + \frac{da_2}{dx} = \frac{\frac{2a_1}{cos(\theta)}\tau_{W_1} + B_{01}\tau_{I_1}}{a_1\rho_1 g\left(B_{0_1} + \frac{S}{2}a_1\right)} + E_{l_1} \tag{4.16}$$

$$\frac{u_2}{g}\frac{du_2}{dx} + (1-\epsilon)\frac{da_1}{dx} + \frac{da_2}{dx} = \frac{\frac{2a_2}{cos(\theta)}\tau_{W_2} + B_{0_2}\tau_B + (B_{0_2} + Sa_2)\tau_{I_2}}{a_2\rho_2 g\left(B_{0_2} + \frac{S}{2}a_2\right)} + E_{l_2} \tag{4.17}$$

Where u is velocity, a is layer height, τ is shear stress, P is wetted perimeter of the layer, E_l is the entrance loses, S is the combined side slopes, ρ is the fluid density, B_0 is the base width of the fluid layer at its lowest point, and θ is the angle of the wall with respect to the horizontal. Here the subscripts 1 and 2 denote upper and lower layer, W denotes the wall location, I denotes the interface location, and B denotes the bottom of the channel.

These layers of flow can also be expressed using the continuity condition given the assumption that there is one set discharge for each layer over the cross section of the channel. This assumption's expression and implication on continuity can be seen in Eq. 4.18.

$$\frac{dQ_i}{dx_i} = 0 = U_i\frac{da_i}{dx_i} + a_i\frac{dU_i}{dx_i} \tag{4.18}$$

Where Q is the set discharge given the layer, U is the velocity, and a is the respective layer height given a specific location.

With the linear momentum and continuity equations, the gradually varying flow can be integrated over the length of the channel. These equations still need to resolve the overall shear stress to be solvable. Thus, to determine the shear stresses for each of the section edges, this study uses Manning's equation for shear stress (see Eq. 4.19).

$$\tau = \frac{f}{8}\rho_i|U_i|U_i \tag{4.19}$$

Where the f value is calculated using Eq. 4.20, for English units, those readily available for the GSL system.

$$f_i = \frac{3.62g}{R_{Hi}^{1/3}} n_i^2 = \frac{116.7}{R_{Hi}^{1/3}} n_i^2 \tag{4.20}$$

Where R_H is the hydraulic radius (see Eq. 4.21), and n_i is the manning friction factor for the given interface.

$$R_{Hi} = \frac{A_i}{P_i} \tag{4.21}$$

Where P_i is the wetted perimeter and A_i is the area of the cross-section. With this understanding the final portion of this equation is to know what the respective hydraulic radius is for each of the subsections inside the single cross-section. Each layer cross-section can be subdivided into a surface (A_s), bottom (A_b), and wall (A_w) flow regions shown in Fig. 4.7.

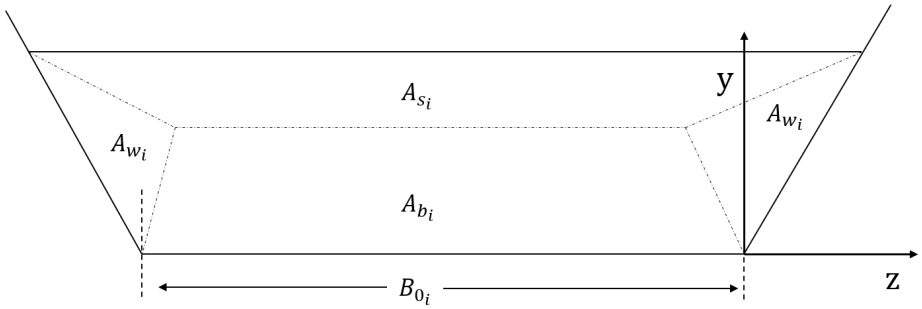

Fig. 4.7: Subdivisions of single flow layer cross-section for respective shear considerations.

From this the total area of the cross-section must be equal to the sum of the individual parts (Eq. 4.22) where shear forces influence the flow.

$$A_i = A_{s_i} + A_{b_i} + A_{w_i} \tag{4.22}$$

The hydraulic radius is solved using Manning's equation (see Eq. 4.23b).

$$Q = UA = \left(\frac{1.49}{n}\right) A\, R_H^{2/3}\, S_0^{1/2} \tag{4.23a}$$

$$U = \left(\frac{1.49}{n}\right) R_H^{2/3}\, S_0^{1/2} \tag{4.23b}$$

Where U is the relative velocity of the flow, R_H is the hydraulic radius, and S_0 is the slope of the channel. Using these equations one can solve for the hydraulic radius of each sub section of the flow layer. Starting with the upper layer, the influences of surface shear are ignored due to the free surface, instead the walls and lower surface shear are considered. The total area of the upper layer is shown in Eq 4.24a.

$$A_1 = a_1 \left[B_{0_1} + \frac{a_1 S}{2} \right] = A_I + A_W = P_{I_1} R_{I_1} + P_{W_1} R_{W_1} \tag{4.24a}$$

$$P_{I_1} = B_{0_1} \tag{4.24b}$$

$$P_{W_1} = \frac{2a_1}{cos(\theta)} \tag{4.24c}$$

Where the subscript W and I are the wall and interface locations respectively. Using Manning's equation (Eq.4.23b) one can solve for the ratio expression for the hydraulic radii in the upper layer.

$$\frac{R_{I_1}}{R_{W_1}} = \left(\frac{n_{I_1}}{n_{W_1}} \right)^{3/2} \left(\frac{u_1 - u_2}{u_1} \right)^{3/2} \tag{4.25a}$$

$$R_{I_1} = R_{W_1} \left(\frac{n_{I_1}}{n_{W_1}} \right)^{3/2} \left(\frac{u_1 - u_2}{u_1} \right)^{3/2} \tag{4.25b}$$

$$R_{W_1} = R_{I_1} \left(\frac{n_{I_1}}{n_{W_1}} \right)^{-3/2} \left(\frac{u_1 - u_2}{u_1} \right)^{-3/2} \tag{4.25c}$$

Substituting these expressions into Eq. 4.24a for the wall and interface radius, the expression of each radii can be found as follows.

$$R_{W_1} = \frac{a_1 \left[B_{0_1} + \frac{a_1 S}{2} \right]}{\frac{2a_1}{cos(\theta)} + B_{0_1} \left[\left(\frac{n_{I_1}}{n_{W_1}} \right)^{3/2} \left(\frac{u_1 - u_2}{u_1} \right)^{3/2} \right]} \tag{4.26a}$$

$$R_{I_1} = \frac{a_1 \left[B_{0_1} + \frac{a_1 S}{2} \right]}{B_{0_1} + \frac{2a_1}{cos(\theta)} \left[\left(\frac{n_{I_1}}{n_{W_1}} \right)^{-3/2} \left(\frac{u_1 - u_2}{u_1} \right)^{-3/2} \right]} \tag{4.26b}$$

Using these solved radii the expected shear stresses can be expressed on the upper region due to the wall and interface locations as shown in eqs. 4.27b,4.27a

$$\tau_{W_1} = \frac{116.7 \rho_1 \eta_W^2}{8 R_{W_1}^{1/3}} \left[\, |u_1| \, u_1 \right] \tag{4.27a}$$

$$\tau_{I_1} = \frac{116.7 \rho_1 \eta_I^2}{8 R_{I_1}^{1/3}} \left[\, |u_1 - u_2| \, (u_1 - u_2) \right] \tag{4.27b}$$

Transitioning to the lower layer flow, shear influences due to the top surface must be considered, it being the interface location between the upper and lower regions, yielding the following for the lower area consideration.

$$A_2 = a_2\left[B_{0_2} + \frac{a_2 S}{2}\right] = A_{I_2} + A_{W_2} + A_{B_2} = P_{I_2}R_{I_2} + P_{W_2}R_{W_2} + P_{B_2}R_{B_2} \tag{4.28a}$$

$$P_{I_2} = B_{0_1} \tag{4.28b}$$

$$P_{W_2} = \frac{2a_2}{cos(\theta)} \tag{4.28c}$$

$$P_{B_2} = B_{0_2} \tag{4.28d}$$

Using the same methods as the upper layer, the hydraulic radii ratios are solved and shown below.

$$\frac{R_{I_2}}{R_{W_2}} = \left(\frac{n_{I_2}}{n_{W_2}}\right)^{3/2}\left(\frac{u_1 - u_2}{u_1}\right)^{3/2} \tag{4.29a}$$

$$\frac{R_{I_2}}{R_{B_2}} = \left(\frac{n_{I_2}}{n_{B_2}}\right)^{3/2}\left(\frac{u_1 - u_2}{u_1}\right)^{3/2} \tag{4.29b}$$

$$\frac{R_{B_2}}{R_{W_2}} = \left(\frac{n_{B_2}}{n_{W_2}}\right)^{3/2} \tag{4.29c}$$

Substituting these ratios into Eq. 4.28a the following hydraulic radii can be solved.

$$R_{W_2} = \frac{a_2\left[B_{0_2} + \frac{a_2 S}{2}\right]}{\frac{2a_2}{cos(\theta)} + [B_{0_2} + a_2 S]\left[\left(\frac{n_{I_2}}{n_{W_2}}\right)^{3/2}\left(\frac{u_1 - u_2}{u_2}\right)^{3/2}\right] + B_{0_2}\left[\left(\frac{n_{B_2}}{n_{W_2}}\right)^{3/2}\right]} \tag{4.30a}$$

$$R_{I_2} = \frac{a_2\left[B_{0_2} + \frac{a_2 S}{2}\right]}{\frac{2a_2}{cos(\theta)}\left[\left(\frac{n_{W_2}}{n_{I_2}}\right)^{3/2}\left(\frac{u_2}{u_1 - u_2}\right)^{3/2}\right] + [B_{0_2} + a_2 S] + B_{0_2}\left[\left(\frac{n_{B_2}}{n_{I_2}}\right)^{3/2}\left(\frac{u_2}{u_1 - u_2}\right)^{3/2}\right]} \tag{4.30b}$$

$$R_{B_2} = \frac{a_2\left[B_{0_2} + \frac{a_2 S}{2}\right]}{\frac{2a_2}{cos(\theta)}\left[\left(\frac{n_{W_2}}{n_{B_2}}\right)^{3/2}\right] + [B_{0_2} + a_2 S]\left[\left(\frac{n_{W_2}}{n_{B_2}}\right)^{3/2}\left(\frac{u_1 - u_2}{u_2}\right)^{3/2}\right] + B_{0_2}} \tag{4.30c}$$

With these radii the expected shear stresses on the lower region due to the interface, wall, and bottom locations can be expressed as shown in Eqs. 4.31a, 4.31b, 4.31c

$$\tau_{I_2} = \frac{116.7\rho_2\eta_I^2}{8R_{I_2}^{1/3}} \left[|u_1 - u_2| \, (u_1 - u_2) \right] \tag{4.31a}$$

$$\tau_{W_2} = \frac{116.7\rho_2\eta_W^2}{8R_{W_2}^{1/3}} \left[|u_2| \, u_2 \right] \tag{4.31b}$$

$$\tau_{B_2} = \frac{116.7\rho_2\eta_B^2}{8R_{B_2}^{1/3}} \left[|u_2| \, u_2 \right] \tag{4.31c}$$

Solving these hydraulic radii allows for the expected shear terms to be calculated for Eqs. 4.16, 4.17. With these terms solved for, there needs to be boundary condition closure of this model. Boundary conditions are enforced using energy equations. Specifically, using Bernoulli's equation (Eq. 4.32) to relate the energy from the larger lake bodies to the expected energy at the entrance and exit of the NB.

$$P_1 + \rho\frac{v_1^2}{2} + \rho g h_1 = P_2 + \rho\frac{v_2^2}{2} + \rho g h_2 \tag{4.32}$$

Where P is pressure, V is velocity, g is gravitational acceleration, ρ is density of the fluid, and h is the water height from a given datum location. Here the subscript 1 and 2 are for two different locations in the same water flow, not denoting different layers of flow like in the rest of this study.

Using Eq. 4.32, three different location pairs are considered; 1) between the south lake arm and the southern end to the NB for the upper layer flow, 2) between the north lake arm and the upper layer flow at the northern end of the NB, and 3) between the north lake arm and the lower layer on the northern side of the NB. For all of these cases the lake elevation is taken from a datum placed at the height of the northern control berm placed in the NB channel. Where as built conditions of the control berm are set at an elevation of 4183 ft (1275 m) from the National Geodetic Vertical Datum (NGVD).

Taking eq 4.32 and applying it to case one listed above yields the following energy boundary condition.

$$H_S = a_{1S} + a_{2S} + \frac{u_{1S}^2}{2g} = a_{1S} + a_{2S} + \frac{1}{2g}\left(\frac{Q_1}{a_{1S}\left[B_{0_1} + \frac{a_{1S}S}{2}\right]}\right)^2 \tag{4.33}$$

Where H_s is the WSE of the southern arm from the prescribed berm datum, a is the respective layer heights, u is the layer velocity,S is the combined slope of the two channel sides, and s subscript denotes the southern end of the NB. Considering the upper layer on the northern end of the NB, the conditions are set between an infinitesimally thing layer that forms beyond the breach before the southern and northern water start to mix. Here the velocity of the upper layer is near zero and rests at the height of the northern arm of the lake. Relating these positions yields the following boundary condition.

$$H_N = a_{1N} + a_{2N} + \frac{u_{1N}^2}{2g} - h_x \tag{4.34}$$

Where H_N is the WSE of the northern arm from the prescribed berm datum, h_x is the exit losses, and N subscript denotes the northern end of the NB. The energy between the northern arm and the lower layer entrance on the northern end of the NB is shown in Eq. 4.35. Equations 4.33, 4.34, 4.35 ensure that lake boundary conditions can be enforced when solving the system.

$$H_N = \frac{\rho_1}{\rho_2}a_{1N} + a_{2N} + \frac{u_{2N}^2}{2g} = \frac{\rho_1}{\rho_2}a_{1N} + a_{2N} + \frac{1}{2g}\left(\frac{Q_2}{a_{2N}\left[B_{0_2} + \frac{a_{2N}S}{2}\right]}\right)^2 \tag{4.35}$$

Using the equations found in this section, one can solve for the two layer heights throughout the NB using the SSWEM. The procedure of computation is dictated by a specific flow case. This will be done by starting with initial conditions for the GSL namely, WSE for the two arms, and their respective densities. For this study, development is conducted for the uni-directional SN and NS cases; leaving bi-directional flow to future

work. Solving for the uni-directional cases, each flow has unique methods to the initial set up of the problem, but the solving method is the same.

The first step of modeling the change in layer height given lake conditions using Eqs. 4.16,4.17,4.32 is to assess the change in velocity expressed in terms of the layer height changes. This is done by rearranging the continuity equation to yield Eq. 4.36.

$$\frac{dU_i}{dx_i} = \frac{-U_i}{a_i}\frac{da_i}{dx_i} \tag{4.36}$$

Using this one can rewrite the upper and lower layer linear momentum equations using Eq. 4.36 to eliminate the need for change in velocity considerations and only have layer height dependencies.

$$\frac{-U_1^2}{ga_1}\frac{da_1}{dx} + \frac{da_1}{dx} + \frac{da_2}{dx} = T_1 \tag{4.37}$$

$$\frac{-U_2^2}{ga_2}\frac{da_2}{dx} + (1-\epsilon)\frac{da_1}{dx} + \frac{da_2}{dx} = T_2 \tag{4.38}$$

Where T is the representation of the shear forces for a given layer. This reduction is used to improve the readability of the overall solution, though does not add additional information to the solver.

$$T_1 = \frac{\frac{2a_1}{cos(\theta)}\tau_{W_1} + B_{01}\tau_{I_1}}{a_1\rho_1 g\left(B_{0_1} + \frac{S}{2}a_1\right)} + E_{l_1} \tag{4.39}$$

$$T_2 = \frac{\frac{2a_2}{cos(\theta)}\tau_{W_2} + B_{0_2}\tau_B + (B_{0_2} + Sa_2)\tau_{I_2}}{a_2\rho_2 g\left(B_{0_2} + \frac{S}{2}a_2\right)} + E_{l_2} \tag{4.40}$$

This study develops the uni-directional flow cases, and leaves bi-directional flow cases for future development. The solving procedure of the bi-directional would be similar to that of the uni-directional flow cases. The solving procedure of SSWEM is: assess a maximum discharge allowed for each layer of flow, use the energy boundary conditions to calculate the entry heights for each layer, integrate the change in layer height over the domain using a predictor corrector explicit stepping, check how the solved discharges compare to the known

lake conditions given, iterate on the solution if necessary to finalize expected discharges, and then repeat the process for each flow instance desired.

Uni-directional SN Flow Case

Therefore, using Eq. 4.37 the uni-directional SN case can be determined by setting $\frac{da_2}{dx} = 0$. The maximum discharge possible can be determined by the maximum cross-sectional area, and the available velocity head between the two sides of the lake. The maximum SN discharge is solved using Eq. 4.41

$$Q_{1Max} = \sqrt{2 * g * (H_S - H_N)} * \left(H_S * \left(B0_1 + \frac{H_S * S}{2} \right) \right) \tag{4.41}$$

Equation 4.41 sets the max discharge and the minimum discharge is set to 0 cms (0 cfs). The tested discharge is the average of these discharge bounds. The beginning layer height is determined using the tested discharge and the energy boundary condition in Eq. 4.33, setting $a_{2s} = 0$. The beginning layer height is iteratively solved for due to a_{1s} not being cleanly separable.

Using this beginning layer height the system is integrated across the channel length using an explicit predictor corrector method; the runge-kutta 2 method. This stepping scheme helps to smooth the transition between one cell to the next and help reduce instabilities in the stepping procedure.

After the stepping routine reaches the north exit, the layer height is checked against the required energy condition, Eq. 4.34 setting the lower layer height to zero. This yields a requirement that the upper layer height at the north end (a_{1N}) of the breach must satisfy the following Eq. 4.42.

$$a_{1N} \geq H_N * \left(\frac{\rho_2}{\rho_1} \right) \tag{4.42}$$

This condition is checked to assess the solved height relative to expected conditions. If the height is too high the minimum discharge is set to the tested discharge, and if the height is too low the maximum discharge is set to the tested discharge. Either way the system iterates until the simulated height is within a desired threshold of the condition set by Eq. 4.42.

Another essential parameter to ensure throughout solving the domain is that the upper layer flow remains supercritical. Where the critical conditions requires that the flow layer's densimetric Froude number (Eq. 4.43) does not fall below 1. If the flow layer becomes sub critical a uni-directional flow case is not possible. To check this method the layer velocity is compared to the flow requirement (Eq. 4.44).

$$Fr_i = \frac{U_i}{\sqrt{\epsilon * g * a_i}} \tag{4.43}$$

$$U_i \geq \sqrt{\epsilon * g * a_i} \tag{4.44}$$

Where Fr is the densimetric Froude number, ϵ is the density, U is the layer velocity, a is the layer height, and g is gravitational acceleration. If this number does become subcritical the assumption of uni-directional flow is violated and then the existence of two layers is known and the solver must account for both layers of flow.

Uni-directional NS Flow Case

Solving for the uni-directional NS flow case is remarkably similar to the methodology of the USN case. In this case the integration of the domain is done from north to south instead of south to north. In this case $\frac{da_1}{dx} = 0$ simplifies the boundary conditions and the layer integration. Solving for the maximum discharge allowed yields the following equation given the available velocity head between the two arms (Eq.4.45).

$$Q_{2Max} = \sqrt{2 * g * (H_N - H_S)} * \left(H_N * \left(B0_2 + \frac{H_N * S}{2} \right) \right) \tag{4.45}$$

Given a set maximum discharge, the minimum discharge is set to 0, and the test discharge is set to the average of the two discharges. The entrance height of the flow layer can be determined using Eq. 4.35, setting $a_{1n} = 0$, requiring an iteratively solved layer height entering the channel. The layer height requirement at the exit of the channel is that the flow layer height remains greater than the south arm height, Eq. 4.46.

$$a_{2S} \geq H_S \tag{4.46}$$

Using the same method as before, the flow layer is iterated across the domain. If the height is below the required condition (Eq. 4.46) the maximum discharge is set to the test discharge. Whereas if the height satisfies this condition but the comparative height to the south arm water height is not within the desired threshold the minimum discharge is set to the test discharge. Similar to the uni-directional SN case, if the flow falls below the critical densimetric Froude number the single layer flow case is no possible and a bi-directional flow case must be assessed. The discharge is iteratively solved for until the height requirement fall within the required threshold.

4.2 Results

4.2.1 Generalized Area Based Index Model

The purpose of GABI's development is to specifically handle cases for which there is not enough data to use ML modeling techniques. These cases stem from changes to the berm elevation in February and July of 2023 to a 4 ft and 9 ft berm height from the channel floor. Though the modeling technique is fundamentally based and less glamorous then more complicated modeling techniques, it still shows promise in being able to predict where other models are unable to.

Through the development of GABI, the model shows a bias in both discharge measurements. In order to correct for this bias, GABI is tested on the 0 ft berm cases and the overall discharge is compared to USGS predicted discharge. Using this data a bias shift

is calculated to ensure GABI predicts within 5 percent of the total discharge. From this analysis the correction factors to the GABI predicted discharge are shown in Eqs. 4.47, 4.48.

$$Q_{1_{adj}} = Q_1 * 1.25 \tag{4.47}$$

$$Q_{2_{adj}} = Q_2 * 0.70 \tag{4.48}$$

The use of a correction factor is not uncommon in models due to the limitations imposed due to the assumptions used. GABI assumes there exists uniform velocity across the total cross section in each cell subdivision. This assumption excludes the effects of wall shear forces leading to frictional losses in each cell. These missing forces account for much of the disparity seen between USGS predictions and GABI predictions. To further refine this code one could create a correction factor for each individual cell. Basing the bias of the cells on the location in the water column height, proximity to walls, and an expected influence on subcell velocities. This kind of correction would be more robust, but is outside the scope of this study. Using all USGS monthly measurements, both the raw and adjusted outputs of GABI are used to approximate the discharges, the results are shown in Fig. 4.8, with full performance listed in Table 4.2

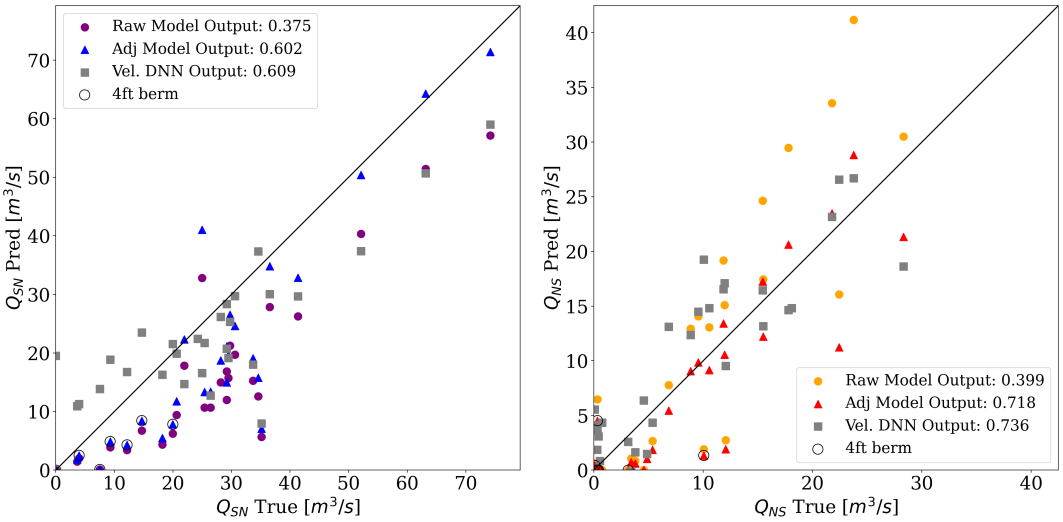

Fig. 4.8: Performance of GABI model compared to USGS monthly recorded measurements both with raw output and adjusted output, showing an overall improvement using a bias shifted discharge prediction.

Here it is clear to see that GABI shows a clean linear trend compared to USGS measurements. This trend is what led to the understanding that a bias shift is necessary to correct for missing information. Using the bias improves the SN prediction NSE value by 0.227, and the NS value by 0.319, both substantial improvements from the base development. Looking at the SN discharge there is still a tendency to under predict even with the bias correction. Whereas the NS discharge plot shows an improved distribution around the ideal prediction line.

Table 4.2: Statistical performance values of GABI compared to USGS monthly discharge measurements

Model Type	Flow Dir.	NSE	RMSE	CC	PBIAS
GABI (raw)	SN	0.375	0.791	0.904	40.8591
	NS	0.399	0.7751	0.889	-15.04
GABI (adj)	SN	0.602	0.6307	0.904	26.07
	NS	0.718	0.531	0.889	19.472
Vel. DNN	SN	0.6094	0.625	0.820	12.787
	NS	0.736	0.514	0.879	-14.0983

Adjusting the discharge output from GABI allows for increase performance and the ability to compensate for some of the idealized assumptions made during development.

Table 4.3: Model performance compared using the bias adjusted GABI predictions and monthly USGS measurements

	SN Flow [cms]			NS Flow [cms]		
Date	USGS	GABI	% Error	USGS	GABI	% Error
Jul-2018	20.620	11.320	-45.110	28.320	21.610	-23.670
Aug-2018	29.170	16.060	-44.940	15.430	16.620	7.700
Sep-2018	24.230	0.000	-100.000	18.100	25.230	39.490
Apr-2019	52.110	50.230	-3.590	4.560	0.000	-100.000
May-2019	74.200	71.740	-3.310	0.550	0.000	-100.000
Jun-2019	63.140	65.260	3.390	5.350	1.500	-72.040
Jul-2019	41.350	35.200	-14.860	15.490	12.170	-21.410
Aug-2019	24.980	38.610	54.590	22.420	12.820	-42.840
Oct-2019	18.180	5.400	-70.320	23.760	28.870	21.550
Jun-2020	35.130	6.630	-81.120	21.770	23.810	9.350

Continued on next page

Table 4.3 – continued from previous page

	SN Flow [cms]			NS Flow [cms]		
Date	USGS	GABI	% Error	USGS	GABI	% Error
Jul-2020	25.400	13.670	-46.180	17.780	20.470	15.090
Oct-2020	26.420	12.340	-53.270	11.870	14.010	18.090
Dec-2020	29.170	22.050	-24.400	9.540	9.200	-3.590
Apr-2021	36.530	34.780	-4.790	6.850	5.440	-20.650
May-2021	29.730	24.740	-16.790	11.950	11.500	-3.750
Jun-2021	30.600	24.090	-21.240	10.560	9.500	-10.030
Jul-2021	28.180	18.560	-34.130	8.830	9.050	2.470
Sep-2021	21.940	23.280	6.100	12.060	1.330	-88.990
Nov-2021	34.540	15.370	-55.520	3.400	0.930	-72.620
May-2022	33.690	18.820	-44.130	3.800	0.700	-81.540
Jun-2022	29.450	19.450	-33.950	4.870	1.100	-77.500
Jul-2022	19.990	7.820	-60.890	3.140	0.000	-100.000
Aug-2022	14.700	8.450	-42.470	0.310	0.000	-100.000
Aug-2022	7.540	0.110	-98.520	0.340	4.530	1243.920
Aug-2022	9.290	4.880	-47.420	0.420	0.000	-100.000
Sep-2022	3.990	2.380	-40.450	0.085	0.510	500.560
Sep-2022	0.000	0.000	0.000	10.050	1.340	-86.670
Nov-2022	3.630	1.850	-49.080	0.740	0.000	-100.000

4.2.2 Steady Shallow Water Exchange-flow Model

The primary results of this study are the development of SSWEM using a trapezoidal cross-section, shown in the methodology section of this document. The focus of this study is to begin the development of this model and getting it to perform as expected. For this study two main cases are focused on, the uni-directional SN case, and the uni-directional NS case. Each of these flow cases allow for a simplified approach to solving the layer propagation

through the channel not having to consider the influence of the second layer. The results of the initial stages of the solver development are shown in Figs. 4.9, 4.10. Using SSWEM, additional parameters are able to be monitored such as the rate of change across the channel and the layer velocity compared to the critical value. The additions are shown in the figure representations, where the original model only provides a final discharge approximation.

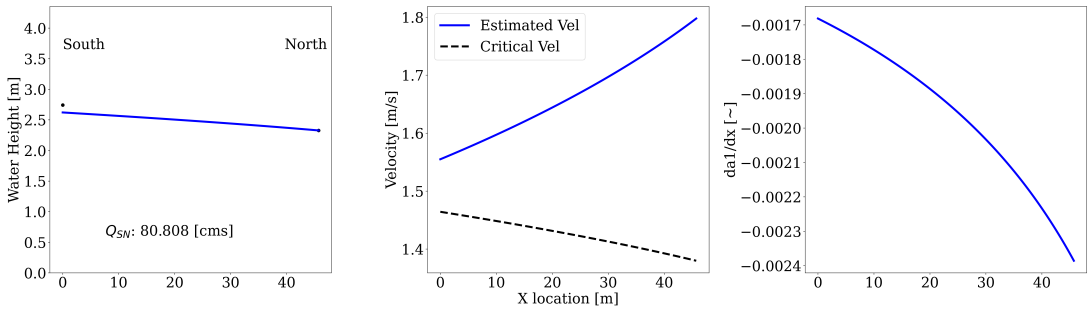

Fig. 4.9: Results of discharge simulation of uni-directional SN flow, including velocity tracking, and rate of change in layer height displayed

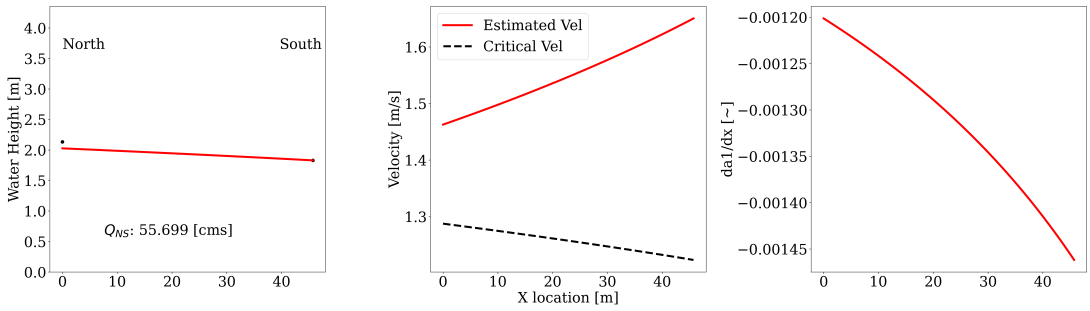

Fig. 4.10: Results of discharge simulation of uni-directional NS flow, including velocity tracking, and rate of change in layer height displayed

The inclusion of additional data monitoring allows for increased visualization of the physics compared to the original model developed by Holley and Waddell. Instead of creating a black box model yielding only discharge values, the newer version allows one to see if and where a flow may become subcritical. This understanding gives a user a better

intuition where a trapped flow layer may exist, consistent with the arrested wedge regime seen in the modeling of Holley and Waddell [1].

From the results shown in both Fig. 4.9 and Fig. 4.10 it is clear to see both layers remain super critical throughout the channel length. It is interesting to note the rate of change in the layer shows a non-linear trend as the model progresses down the channel. This result is due to the shear forces in the model. The shear forces scale with velocity squared, instead of scaling linearly. This being true, the overall change in layer height over the channel though still exhibits a linear trend consistent with what is expected from a gradually varying flow.

The initial development of SSWEM shows promise in being able to increase access and interpretability to information compared to previous models used. There are still additional modeling efforts to fully develop SSWEM, the first of which being a procedure to model the bi-directional flow through the breach. Current instabilities that exists in the 1D model are demonstrated at higher discharge values. An example of this is shown in Fig. 4.11, demonstrated using a uni-directional SN flow case with higher discharge value.

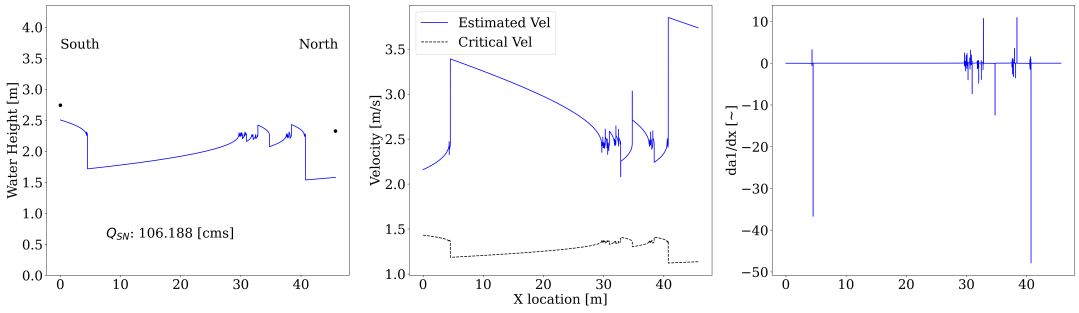

Fig. 4.11: Example of instabilities in solver code brought by higher discharge values, example case shown using uni-directional SN flow case.

Instabilities within SSWEM become more noticeable as the discharge, and by extension velocity, increases. This in turn increases the expected shear forces, which influences the rate at which the layer height change. The reduced layer height increases the velocity from the continuity constraint set by Eq. 4.18. Since an explicit solving technique (Eq. 4.49) is

used, the solving method is constrained by a stability condition. As the velocity is increased the required step size becomes smaller and smaller. Therefore, to fix this issue one would need to increase the resolution of the domain to allow for smaller steps to be taken. Another method to fix this stepping instability is to utilize an implicit solver technique (Eq. 4.50). Implicit solvers are unconditionally stable, meaning you can take larger step sizes while generating a correct rate of change to the next position. The difference between an implicit and explicit solver is in what variables the rate of change is dependent on.

$$a_{j+1} = a_j + \frac{da_j}{dx} * \Delta x \tag{4.49}$$

$$a_{j+1} = a_j + \frac{da_{j+1}}{dx} * \Delta x \tag{4.50}$$

Where a is the layer height, and j in this case denotes the current and next cell solved for in the flow layer. In these equations it is seen that the implicit solve rate of change is dependent on the next step's variables, while the explicit is only dependent on the current step. The difficulty for this specific application using an implicit solved method comes in the dependencies of shear on velocity and velocities dependence on layer height. Due to this fact, the implicit stepping technique is difficult to implement because a_{j+1} and a_j are not cleanly separable. With the difficulty of implicit stepping, this study only implements explicit stepping.

4.2.3 Future SSWEM solver developments

Given the results of SSWEM, it is clear the new implementation has further development required. The main issue with the new implementation is the missing consideration of the new berm geometry. With the dynamic changes made to the control berm structure, it is necessary for modeling to account for such changes.

The inclusion of the control berm in the solving method can be achieved by breaking the channel into separate sections. Solving for the flow layers in each channel subsection iteratively. The subsections are broken into three segments: the north end of the NB to the

berm, flow across the berm, and then the berm to the south end of the NB shown in Fig. 4.12.

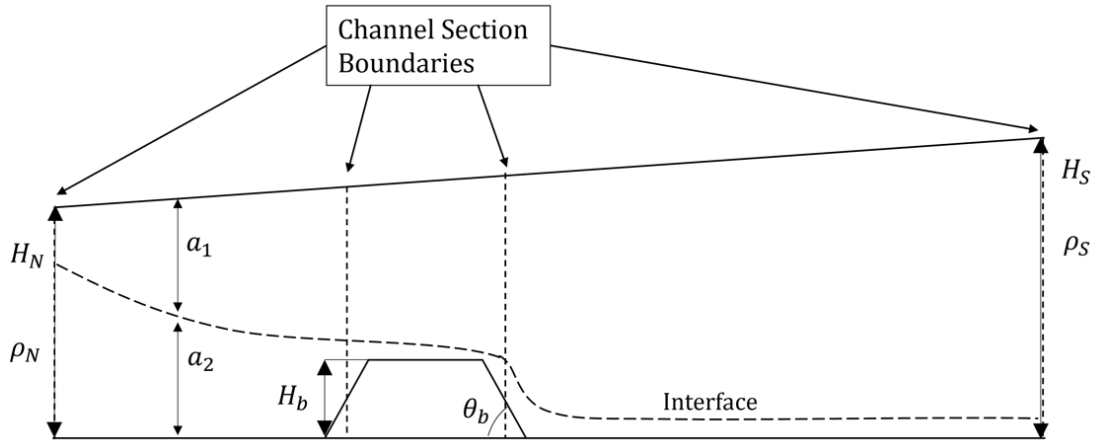

Fig. 4.12: Section solve method for 1D model implementation to include effects of a control berm to the flow dynamics.

Solving the NB flows using this subsection method allows the berm to be considered as a flow constriction between the north and south ends of the channel. Due to this constriction, the flow will encounter another loss term distributed over the length of the berm surface. Therefore, to consider the berm geometry one needs to solve two additional energy equations at the beginning and end of the breach. This can be done using the same energy equations found in the berm-less SSWEM solver shown in this study. Therefore, when integrating over the length of the berm the upper and lower layer flows will be governed by equations similar to those shown in Eqs. 4.51, 4.52.

$$\frac{-U_1^2}{ga_1}\frac{da_1}{dx} + \frac{da_1}{dx} + \frac{da_2}{dx} = T_1^* \tag{4.51}$$

$$\frac{-U_2^2}{ga_2}\frac{da_2}{dx} + (1-\epsilon)\frac{da_1}{dx} + \frac{da_2}{dx} = T_2^* \tag{4.52}$$

Where T^* is the representation of the shear forces for a given layer with the inclusion of berm loss factors. This reduction is used to improve the cleanliness of the overall solution, though does not add additional information to the solver.

$$T_1^* = \frac{\frac{2a_1}{cos(\theta)}\tau_{W_1} + B_{01}\tau_{I_1}}{a_1\rho_1 g\left(B_{0_1} + \frac{S}{2}a_1\right)} + E_{l_1} + E_{b_1} \tag{4.53}$$

$$T_2^* = \frac{\frac{2a_2}{cos(\theta)}\tau_{W_2} + B_{0_2}\tau_B + (B_{0_2} + Sa_2)\tau_{I_2}}{a_2\rho_2 g\left(B_{0_2} + \frac{S}{2}a_2\right)} + E_{l_2} + E_{b_2} \tag{4.54}$$

Where here the E_{b_i} term is the distributed loss factor resulting from the losses due to the berm constricting the flow path. This loss is formulated consistent with flow constrictions following Eq. 4.55

$$E_{b_i} = K_i * \frac{U_i^2}{gL_b} \tag{4.55}$$

Where K is the loss factor given berm geometry, U is the velocity of the layer over the berm, and L_b is the length of the top of the berm. The layer heights will be solved the same way as the general method, though they will have to conform to the additional energy boundary conditions at the start and end of the berm. This solving technique is not currently implemented and is instead left for future development. This study instead provides the developed fundamental understanding required include berm influences on layer height propagation.

CHAPTER 5
DISCUSSION

This chapter will discuss the results of this study an their impact and contributions to ongoing research to understand buoyancy-drive exchange flows of the GSL and predict the associated flow discharges values. The results of this study provide valuable insight to the capabilities of machine learning for flow prediction, and the limitations when considering ML methods for a physically-based system. In addition to this contribution, this research also shows how physically-based numerical models can be utilized to overcome data constraints of machine learning methods when systems do no allow for proper dataset size.

The culmination of this research encompasses a large contribution to increase prediction speed, versatility, and fidelity of GSL models. This research shows when encountering data limitations there are still various ways one can solve a complex system when relying on fundamental principles. It also shows, when given a proper dataset one can utilize advancements in computation methods to generate reliable discharge predictions to inform future management projects for the GSL faster than more expensive CFD simulations.

5.1 Machine Learning Models

The results of this study demonstrate that complex buoyancy-driven exchange flow in the GSL NB can be modeled using machine learning methods. The ML methods of this study include linear regression, random forests, support vector regression, and deep neural networks. The training dataset comes from USGS data collection sites located around the breach. These models enhance the ability to predict GSL flows, while leveraging existing data from USGS monitoring sites.

However, with four different monitoring sites, there can be issues with all monitoring sites consistently giving the full dataset required for flow prediction. There are multiple instances were one or more sensors may fail to record their designated data, creating gaps in

the data and resulting in a reduced dataset. This limitation is an inherent challenge when working with real world systems, were consistent monitoring is not always guaranteed.

This limitation has a profound impact on the dataset size available for machine learning models. Using linear interpolation and knowing the physical trends at the GSL, data gaps can be filled, though this is done conservatively. To further improve the dataset, increased consistency in monitoring sites and reliability of measurement methods are essential. Additionally, improvement could be achieved by increasing the number of data sampling sites to create redundancies, ensuring consistent data collected. Despite these known data limitations, machine learning methods demonstrate their capability for flow prediction using multiple dataset configurations.

Through the development of ML models using the hydrological dataset, random forest methods performed best based on the performance metrics of this study. Each model exhibited issues at the high and low discharge values, especially at the zero discharge in SN flow. This problem is likely caused by the case disparity between bi-directional and uni-directional flow cases. The focus of the machine learning methods is on the bi-directional flow cases because these are the most predominant case shown in the dataset. To minimize this class disparity, one could decrease the dataset size further to create closer class distributions. However, doing so would further limit the machine learning methods due to a decreased number of data instances.

While the developed models perform worse at extreme discharge locations, removing these cases from the dataset typically improves the general performance of the models. Among the models, the RF model performs best, with an SN NSE value of over 0.90, with an NS NSE of close to 0.9. This result shows the ML models are well behaved for the general bi-directional flow cases, though have lower performance in the uni-directional cases. When comparing the developed models to the monthly USGS measurements each of the methods performance decreases. This decline is due to behaviors not seen in the training set being present in the monthly dataset. The USGS measurements may exhibit different patterns

in the dataset that the ML models are unable to predict. The reduced performance is attributed to out of dataset instances, though SN predictions still do fairly well.

Comparing the linear regression and random forest models to the current 1-D model showed machine learning models perform better. The 1-D model was originally developed for the box culverts placed in the causeway to allow for exchange flow. With the new NB, the original development breaks down, and is unable to handle the larger flow cases. This results in a trend of over-prediction seen in 1-D model results for SN flow, and under-prediction in the NS flows.

Utilizing the flow predictions from the 1-D model, hybrid neural networks were created. This method was expected to increase the fidelity of the machine learning models by including information from a physically-based numerical model. The results indicate the inclusion of this data does not drastically change the model performance given the parallel method, and greatly reduces performance with the series model. A better method to utilize in the future may be PINN models, where the flow equations are incorporated as additional loss factors for gradient formulation in the neural network. This type of network was outside the scope of this project, but may provide a better way to incorporate the desired physical constraints on the network.

Further utilization of the dataset focused on velocity data collected from a single sensor. Given the known limitations of multiple sensors, using a single sensor allows for more consistently available data. This approach increased the instances in the dataset from $\approx$ 24,000 to $\approx$ 52,000 useable instances. The consistent velocity data, increased dimensionality of input, and dataset size improved the performance of the machine learning models. From these results the RF models still outperformed other models and proved to be the best method of prediction.

While these models are capable of performing well, machine learning models are limited by the dataset used for their development. Due to physical changes to the GSL NB structure via the control berm, the as-built dataset used in this study is unable to train models for the current NB structure with a 9 ft berm. The changes in berm height to the 4 ft and 9

ft configurations physically alter the flow dynamics in the NB and there is not enough data to create machine learning models for these flow structures. If machine learning is desired for these structures, additional monitoring must be conducted to produce flow predictions and generate a dataset for model development.

5.2 Physically-Based Numerical Models

From the results of this study, machine learning models are capable of predicting the buoyancy-driven exchange flow at the GSL. However, due to changes in the physical flow structure at the NB there is not enough data to use machine learning models for flow prediction of the 4ft and current 9ft flow cases.

To overcome this issue the physically-based models were derived, allowing for flow prediction based primarily on physical conditions at the NB. The first of these models is GABI, which uses known channel geometry and measured velocity measurements. This method allows for flow prediction in cases where machine learning can not predict like in the presence of a raised control berm. GABI primarily contributes by being able to predict discharge when other methods can not. However, GABI still has its own limitations due to idealized assumptions used for velocity consideration in the discharge solving method. Further development can be conducted to improve GABI's accuracy by localizing bias to individual cells instead of full flow layers.

A more refined approach to solving for discharge based on physical conditions is SSWEM, where gradually varied shallow water equations are derived using the trapezoidal cross-section instead of a box culvert geometry. Where SSWEM is not constrained by a dataset and instead is a physics based model that uses fundamental principles to create the appropriate flow equations.

The contribution of SSWEM, compared to the current 1-D model, is demonstrated in the additional variable monitoring available. The current 1-D model functions as a black-box model, taking in lake conditions and producing discharge approximations. In contrast, SSWEM provides more than just discharge; it includes velocity data, layer height, and rate of change throughout the channel length. This additional data allows users of the model to better understand the flow dynamics and gain deeper insights into the GSL system.

To further enhance the SSWEM, it should be segmented into three subsections totaling the length of the channel. This segmented approach allows for the addition of the berm influence through an additional distributed loss term. Utilization of SSWEM will enable researchers and GSL management to make informed decisions by considering all physical characteristics of the GSL NB channel. Future efforts can either utilize this model or implement additional monitoring to build larger datasets suitable for creating machine learning models to predict flow. Given the current monthly to bi-monthly sampling frequency, there will not be sufficient data to effectively utilize machine learning. Therefore, such endeavors should not be pursued until larger datasets become available.

CHAPTER 6

CONCLUSION

In conclusion, the results of this study represent a significant contribution to the capability to predict buoyancy-driven flows in the Great Salt Lake using both machine learning and physically based modeling methods. Modeling a real-world physical system is inherently complex and challenging, no matter the scale. This study has demonstrated effective modeling of the GSL using multiple methods and has provided valuable insights into the limitations of each method based on data availability and underlying assumptions.

The primary objective of this study evaluate and quantify the efficacy of ML models to predict the NB flows. This study has shown that random forest predictors have the highest performance amongst the ML models used. Each of the developed ML methods represent a contribution to improve rapid prediction and computational cost compared to more extensive CFD simulations. Furthermore, these ML models have shown improved performance over the current 1D modeling system used for predicting exchange flows.

Utilizing the outputs of the current 1D model, hybrid networks were developed to serve as a transfer learning process from physical to data-driven approaches. However, the results of this study indicate that this approach does not yield improvement over a standard ML method. To enhance this approach further, the use of PINNs is necessary to integrate flow equations into the gradient descent method of the model. This understanding emphasizes that incorporating physical parameters into a ML method must extend further than additions to the data, and instead be considered in the model training process.

From the physically based models developed in this study, it was seen that GABI provides a capability to estimate discharge for instances other models can not. Machine learning models struggle to predict scenarios outside the scope of their training set, even more when the physical parameters of the flow system are changed. The lack of discharge data given the changed geometry limits the further development of ML methods. Whereas

GABI requires only velocity considerations and an understanding of the channel cross section to generate predictions. Due to this fact GABI is a significant contribution to the prediction capabilities available at the GSL given a changing berm structure.

In addition to the contributions of GABI, the initial development of a new 1D solving method has demonstrated the capability to predict channel discharges based solely on lake water elevations and corresponding densities. This model enhances the ability to solve 1D gradually varied flow while incorporating the impacts of a changing berm structure. Furthermore, the new 1D model facilitates the availability of flow layer data, thereby enabling deeper insights into the dynamics of flow layers as they move through the NB channel.

In summary, this study not only serves to increase the understanding of the efficacy of ML methods to predict complex flow behavior ,but also highlights considerations regarding the limitations of monitoring sites and models used for prediction. This research clarifies the scenarios where ML methods are suitable for flow prediction and underscores the necessity of employing physically based systems when appropriate. Additionally, this study lays the groundwork for future developments and enhancements to the new 1D modeling method. This study has set the foundation for improvements to the currently implemented 1D solver by considering the trapezoidal cross section and effects of a flow constriction due to a changing berm geometry. These contributions enhance the capabilities of researchers and water management personnel to forecast buoyancy-driven exchange flows in the Great Salt Lake, thereby improving overall system management.

REFERENCES

[1] E. R. Holley and K. M. Waddell, "Stratified flow in great salt lake culvert," *Journal of the Hydraulics Division*, vol. 102, no. 7, pp. 969–985, 1976.

[2] M. Rasmussen, S. Dutta, B. Neilson, and B. Crookston, "Cfd model of the density-driven bidirectional flows through the west crack breach in the great salt lake causeway," *Water*, vol. 13, no. 17, p. 2423, 2021. [Online]. Available: https://doi.org/10.3390/w13172423

[3] M. Freeman, "Flow reversal events and statistical modeling of flow dynamics of hypersaline water across a constructed causeway, great salt lake, utah, usa," Master's thesis, University of Utah, 2014.

[4] HDR, "Annual reports, 2017-2019," received directly from HDR, 2017-2019, https://documents.deq.utah.gov/water-quality/standards-technical-services/gsl-website-docs/uprr-causeway/.

[5] D. Hahl and A. Handy, "Great salt lake, utah: Chemical and physical variations of the brine, 1963-1966," *Utah Geological and Minearlogical Survey Water-Resources Bulletin*, 1969. [Online]. Available: https://pubs.usgs.gov/publication/70048802

[6] USGS, "Usgs water data mapper," https://maps.waterdata.usgs.gov/mapper/index.html, 2023.

[7] C. W. M. Brian L. Loving, Kidd M. Waddell, "Water and salt balance of great salt lake, utah, and simulation of water and salt movement through the causeway, 1987-98," *Water-Resources Investigations Report*, 2000.

[8] C. A. Rumsey, S. A. Hynek, E. larsen, B. M. Crookston, and S. Dutta, "Salt cycling in a terminal lake: Dynamic salt flux at the decadal scale (2010-2022), great salt lake, utah, u.s.a." *Geological Society of America*, 2023, unpublished.

[9] U.S. Geological Survey, "Usgs water data: Great salt lake," 2023, accessed: January 2, 2024. [Online]. Available: https://waterdata.usgs.gov/nwis/uv?site_no=10010025

[10] ——, "Usgs water data: Great salt lake," 2023, accessed: January 2, 2024. [Online]. Available: https://waterdata.usgs.gov/nwis/uv?site_no=10010026

[11] ——, "Usgs water data: Great salt lake," 2023, accessed: January 2, 2024. [Online]. Available: https://waterdata.usgs.gov/nwis/uv?site_no=10010027

[12] S. Dutta, B. M. Crookston, M. Rasmussen, and E. Larsen, "Predicting flow through the causeway of the great salt lake using hydrodynamic simulations and artificial neural networks," Utah State University, Report Paper 679, 2021.

[13] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 1995, pp. 278–282 vol.1.

[14] G. A. Papacharalampous and H. Tyralis, "Evaluation of random forests and prophet for daily streamflow forecasting," *Advances in Geosciences*, vol. 45, pp. 201–208, 2018. [Online]. Available: https://adgeo.copernicus.org/articles/45/201/2018/

[15] S. Desai and T. B. Ouarda, "Regional hydrological frequency analysis at ungauged sites with random forest regression," *Journal of Hydrology*, vol. 594, p. 125861, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169420313226

[16] M. S. Gizaw and T. Y. Gan, "Regional flood frequency analysis using support vector regression under historical and future climate," *Journal of Hydrology*, vol. 538, pp. 387–398, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169416302323

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep 1995. [Online]. Available: https://doi.org/10.1007/BF00994018

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[19] M. Sit, B. Z. Demiray, Z. Xiang, G. J. Ewing, Y. Sermet, and I. Demir, "A comprehensive review of deep learning applications in hydrology and water resources," *Water Science & Technology*, vol. 82, no. 12, p. 2635–2670, 2020.

[20] E. Zhu, Y. Wang, and X. Yuan, "Changes of terrestrial water storage during 1981–2020 over china based on dynamic-machine learning model," *Journal of Hydrology*, vol. 621, p. 129576, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169423005188

[21] R. Gowri, P. Dey, and P. Mujumdar, "A hydro-climatological outlook on the long-term availability of water resources in cauvery river basin," *Water Security*, vol. 14, p. 100102, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468312421000195

[22] E. Kroupi, M. Kesa, V. D. Navarro-Sánchez, S. Saeed, C. Pelloquin, B. Alhaddad, L. Moreno, A. Soria-Frisch, and G. Ruffini, "Deep convolutional neural networks for land-cover classification with Sentinel-2 images," *Journal of Applied Remote Sensing*, vol. 13, no. 2, p. 024525, 2019. [Online]. Available: https://doi.org/10.1117/1.JRS.13.024525

[23] P. Hosseinzadeh, A. Nassar, S. Boubrahimi, and S. Hamdi, "Ml-based streamflow prediction in the upper colorado river basin using climate variables time series data," *Hydrology*, vol. 10, no. 2, p. 29, 2023. [Online]. Available: https://doi.org/10.3390/hydrology10020029

[24] J. Lee, A. Abbas, G. W. McCarty, X. Zhang, S. Lee, and K. Hwa Cho, "Estimation of base and surface flow using deep neural networks and a hydrologic model in two watersheds of the chesapeake bay," *Journal of Hydrology*, vol. 617, p. 128916, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002216942201486X

[25] S. C. Worland, S. Steinschneider, W. Asquith, R. Knight, and M. Wieczorek, "Prediction and inference of flow duration curves using multioutput neural networks," *Water Resources Research*, vol. 55, no. 8, pp. 6850–6868, 2019. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR024463

[26] F. Tut Haklidir and M. Haklidir, "Prediction of reservoir temperatures using hydrogeochemical data, western anatolia geothermal systems (turkey): A machine learning approach," *Natural Resources Research*, vol. 29, pp. 233–2346, 2020.

[27] Keras, "Keras documentation," https://keras.io/, 2023.

[28] M. Ghosh, S. Dutta, and D. Sen, "Sediment flushout from pond and river diversion barrages by gate operation," *Water Resources Management*, September 2014.

[29] F. Chollet, *Deep Learning with Python*. Manning Publishing, 2018.

[30] N. Gupta, "Artificial neural networks," *Network and Complex Systems*, vol. 3, no. 1, 2013.

[31] X. Ying, "An overview of overfitting and its solutions," *Journal of Physics: Conference Series*, vol. 1168, no. 2, 2019.

[32] K. Tsakiri, A. Marsellos, and S. Kapetanakis, "Artificial neural network and multiple linear regression for flood prediction in mohawk river, new york," *Water*, vol. 10, p. 1158, 2018. [Online]. Available: https://doi.org/10.3390/w10091158

[33] N. Gudino-Elizondo, T. Biggs, R. Bingner, E. Langendoen, T. Kretzschmar, E. Taguas, K. Taniguchi-Quan, D. Liden, and Y. Yuan, "Modelling runoff and sediment loads in a developing coastal watershed of the us-mexico border," *Journal of Water*, vol. 11, 2019.

[34] Z. Di, M. Chang, P. Guo, Y. Li, and Y. Chang, "Using real-time data and unsupervised machine learning techniques to study large-scale spatio-temporal characteristics of wastewater discharges and their influence on surface water quality in the yangtze river basin," *Journal of Water*, vol. 11, 2019.

[35] A. Oguz and O. F. Ertugrul, "A survey on applications of machine learning algorithms in water quality assessment and water supply and management," *Water Supply*, vol. 23, no. 2, pp. 895–922, 02 2023. [Online]. Available: https://doi.org/10.2166/ws.2023.033

[36] K. Ng, Y. Huang, C. Koo, K. Chong, A. El-Shafie, and A. Najah Ahmed, "A review of hybrid deep learning applications for streamflow forecasting," *Journal of Hydrology*, vol. 625, p. 130141, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169423010831

[37] K. P. Tripathy and A. K. Mishra, "Deep learning in hydrology and water resources disciplines: concepts, methods, applications, and research directions," *Journal of Hydrology*, p. 130458, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169423014002

[38] P. Tahmasebi and M. Sahimi, "Special issue on machine learning for water resources and subsurface systems," *Advances in Water Resources*, vol. 149, p. 103851, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0309170821000063

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," https://scikit-learn.org/stable/, 2011.

[40] "Tensorflow adam optimizer documentation," https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam.

[41] M. Raissi, P. Perdikaris, and G. 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999118307125

[42] H. Tracy, T. B. Nolan, and F. A. Seaton, "Discharge characteristics of broad-crested weirs," Published Reference Manual, U.S. Department of the Interior, 1957, accessed: 8-20-2023.