# Enhancing Oceanic Variables Forecast in the Santos Channel by Estimating Model Error with Random Forests

**Felipe M. Moreno**[*1] , **Caio F. D. Netto**[1] , **Marcel R. de Barros**[1] , **Jefferson F. Coelho**[1] , **Lucas P. de Freitas**[1] , **Marlon S. Mathias**[2] , **Luiz A. Schiaveto Neto**[1] , **Marcelo Dottori**[3] , **Fábio G. Cozman**[1] , **Anna H. R. Costa**[1] , **Edson S. Gomi**[1] , **Eduardo A. Tannuri**[1]
Center for Artificial Intelligence (C4AI) – University of Sao Paulo, Brazil
[1]Escola Politécnica – University of Sao Paulo, Brazil
[2]Instituto de Estudos Avançados – University of Sao Paulo, Brazil
[3]Instituto Oceanográfico – University of Sao Paulo, Brazil
{felipe.marino.moreno, caio.netto, marcel.barros, jfialho, lfreitasp2001, marlon.mathias, fgcozman, anna.reali, eduat, mdottori}@usp.br

## Abstract

In this work we improve forecasting of Sea Surface Height (SSH) and current velocity (speed and direction) in oceanic scenarios. We do so by resorting to Random Forests so as to predict the error of a numerical forecasting system developed for the Santos Channel in Brazil. We have used the Santos Operational Forecasting System (SOFS) and data collected in situ between the years of 2019 and 2021. In previous studies we have applied similar methods for current velocity in the channel entrance, in this work we expand the application to improve the SHH forecast and include four other stations in the channel. We have obtained an average reduction of 11.9% in forecasting Root-Mean Square Error (RMSE) and 38.7% in bias with our approach. We also obtained an increase of Agreement (IOA) in 10 of the 14 combinations of forecasted variables and stations.

## 1 Introduction

Forecasting of metocean conditions in coastal regions and waterways is an essential task in planning coastal and navigation operations. Forecasts of current and sea surface height (SSH) of water bodies have traditionally been made through numerical models that rely on the solution of simplified Navier-Stokes equations. Those models have inherent errors due to simplifications and uncertainties in parameters and boundary conditions.

An alternative to numerical models is to use machine learning (ML) to infer patterns in previous data measured in the region of interest and thus provide a forecast based only on the interpolation and extrapolation of those patterns observed in the past. However, since ML models only rely on correlations between data, thus ignoring the underlying physics of the problem, they fail if there is a change in the distribution of the data.

A recent and promising line of work consists of combining ML with physics-based models — often referred to as Physics-Informed Machine Learning (PIML). Such an approach aims to take advantage of both the power of pattern recognition given by ML approaches and the power of generalization in unseen scenarios given by the physics-based model.

This work expands on our previous work [Moreno *et al.*, 2022] where PIML was used to correct the error predicted by a numerical model of the speed of water current in a measuring station. Our main contribution here consists of inserting a correction for the direction of the water current and the sea surface height (SSH) predicted by the numerical model into the PIML model. In addition, we expand the corrections to other measurement stations in the Santos-São Vicente-Bertioga Estuarine System region on the Brazilian coast. By producing a direct estimate of the numerical model error, one can correct the model and improve the prediction accuracy.

Section 2 introduces works that also used ML models to correct predictions made by numerical models. Section 3 explains in detail our proposal, while Section 4 describes the experimental setup and the experiments conducted. Section 5 presents the results of the experiments performed and some discussion. Finally, Section 6 presents our conclusions and highlights future work.

## 2 Related Work

PIML is a relatively recent but already vast field of study; review articles such as those by [Willard *et al.*, 2020] and [Kashinath *et al.*, 2021] compile a myriad of applications that already embed ML with physics-based models for different purposes.

Some forecasting applications benefit from using traditional ML techniques to improve the results of physics-based models. For example, [Xu and Valocchi, 2015] use Random Forests (RF) and Support Vector Machine (SVM) to improve

[*]Main and corresponding author. All other authors made significant suggestions and the last five authors oversaw the whole project and secured funding.

the predictive accuracy of groundwater flow numerical models and provide more robust prediction intervals. We now list a few other proposals in the literature that are most relevant to our purposes.

[Eccel *et al.*, 2007] perform a comparison of linear and nonlinear ML models as methods for post-processing the direct outputs of numerical weather forecast models to reduce the biases introduced by a coarse horizontal resolution. The system was used to predict minimum temperatures in a region of the Italian Alps. Artificial Neural Networks (ANN), RF, and a Multi-Linear Model were evaluated, showing similar performance.

[Cho *et al.*, 2020] evaluate the use of RF, SVM, ANN, and a multi-model ensemble to correct a numerical weather prediction model that outputs next-day maximum and minimum air temperatures in Seoul, South Korea. Hence ML is used to mitigate the systematic bias in air temperature forecasting caused by a coarse grid resolution and lack of parameterizations of the numerical model. The study showed that the multi-model ensemble had better generalization performance than the three single ML models.

As shown in the literature, the physical model error is one of the main obstacles to improving the accuracy and reliability in numerical weather and climate prediction. [Bonavita and Laloyaux, 2020] use Multi-Layer Perceptron ANN with three layers to model error estimation and correction in the numerical model temperature and pressure prediction. [Vashani *et al.*, 2010] make a comparative evaluation of different ML post-processing models for numerical prediction of temperature forecast over Iranian territory, concluding that ANN provides the best results. Another type of ANN, Convolutional Neural Networks (CNNs), have been used to reduce temperature forecasting errors in the Scandinavian Peninsula [Isaksson, 2018]. In that work a CNN receives a grid of forecasted values of temperature and other environmental parameters, and produces forecasts with smaller errors than those obtained with a Kalman filter post-processor. CNNs have also been employed by [Chapman *et al.*, 2019] to learn and correct North American atmospheric river forecasts, and by [Scher and Messori, 2018] to predict uncertainty in weather forecast.

Finally, our previous work also applied RF in order to learn the error model of the forecast of sea surface current speed made by a numerical model for the entrance of the Santos Channel, in Brazil [Moreno *et al.*, 2022]. In that work we only corrected the current speed in one station, while in the present paper we improve the PIML model by expanding its application to other stations and other two variables types, current direction and SSH.

## 3 A Proposed SOFS + RF Architecture

We propose a PIML approach to improve the prediction made by the *Santos Operational Forecasting System* (SOFS) [Costa *et al.*, 2020] in the Santos Channel, a key sea region in Brazil. The Santos Channel (see Figure 1) provides access to the Santos Port Complex, the largest port in Latin America with an yearly handling of about 145 million tons of cargo. The channel contains 5 measurement stations maintained by the Marine Pilots, where current and SSH measurements are captured. In addition, there are weather sensors in the region. A numerical model based on physics is implemented within SOFS so as to forecast relevant quantities.

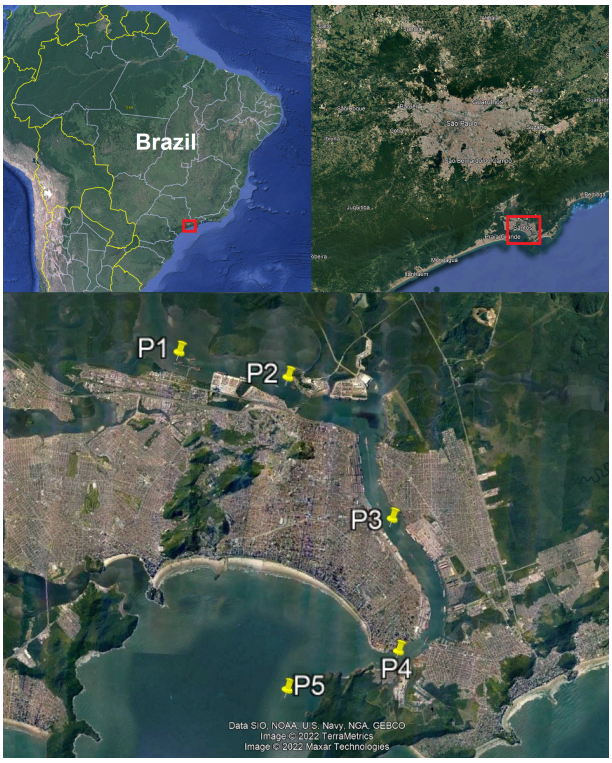


Figure 1: The top images are panoramic views of the Brazilian coast, to locate the Santos Channel. The bottom image shows the Santos Channel, and the five markers indicate the location of measuring stations kept by the Port Authority. Map Source: Google Maps.

Physics-informed techniques in machine learning can incorporate the physics of a domain of interest in different forms, as described for instance by [Willard *et al.*, 2020]. Some approaches incorporate the physics directly into the architecture used to learn various quantities, for example by taking equations to guide the loss function of an ANN during training. Other approaches are inspired by the physical problem so as to guide the design of the architecture, for example by providing the same boundary conditions used for a numerical model solver as inputs of the ML. Another approach is to use a stand alone numerical model and use the ML algorithm to correct its output by either estimating the model error or pondering it with other runs made with perturbed inputs.

Because we have a numerical model already in use for the variables we are interested in, we adopted the third approach by developing an architecture that uses Random Forests (RFs) as an ML model to estimate the SOFS error, which is then corrected in a post-processing phase. RF was chosen for its simplicity, effectiveness and efficiency, also demonstrated in previous work [Eccel *et al.*, 2007]. However, it is worth noting that any other ML model or even an ensemble of ML models could be used for this function, as for instance argued by [Cho *et al.*, 2020].

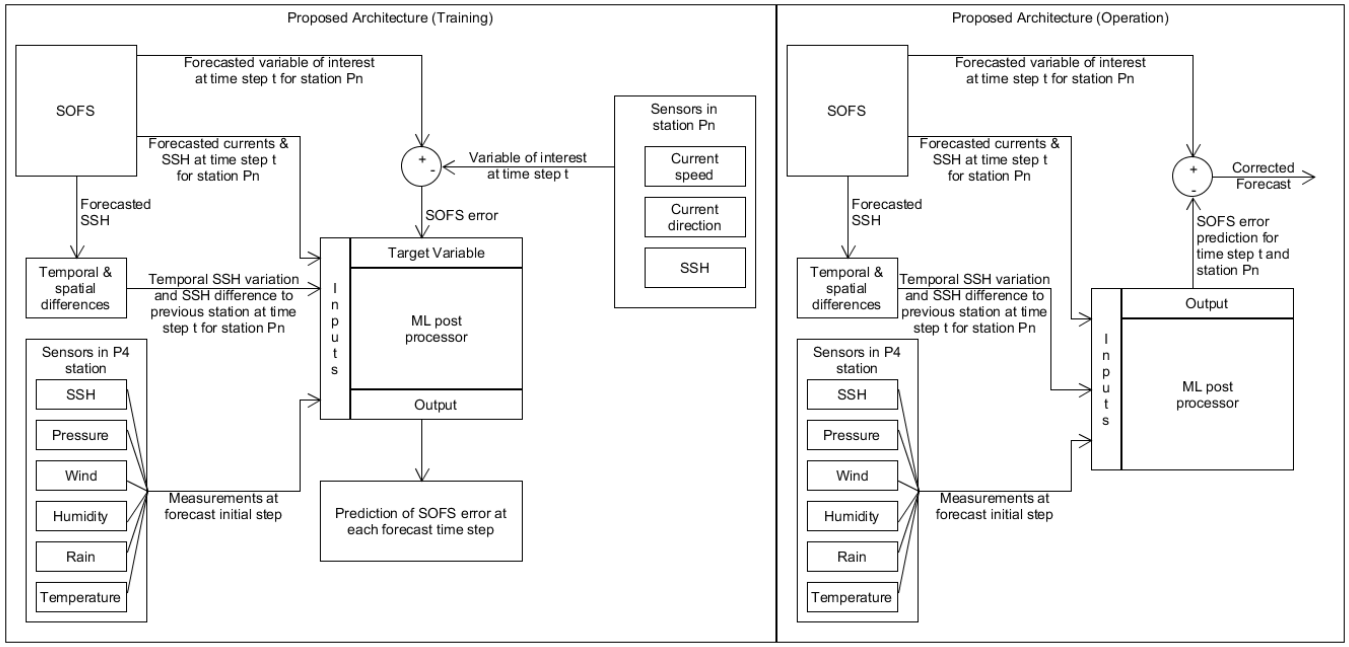


Figure 2: Our proposed Architecture. The training phase is illustrated in the left, where the ML model learns to estimate the SOFS error from the values of its input variables. The operation phase is illustrated in the right, which receives the input variables together with the SOFS estimate and provides the corrected estimate value at the output. There is an ML model trained specifically for each target variable and $P_n$ sensor station.

The target variables estimated by our architecture are the current speed and direction, and SSH at each time step of the SOFS forecast. A distinct RF is trained for each pair of target variable and measurement station $P_n$, with $n \in \{1, 2, 3, 4, 5\}$ (see Figure 1), using the numerical model predictions and sensor data. Once the ML model is trained, the RFs are then used to correct the SOFS prediction error in the operation phase.

An overview of the architecture is depicted in Figure 2. The data used by the PIML architecture comes from the SOFS model and sensors. The SOFS model, the collected data, and the training and operating phases of the PIML architecture are described in the following.

### 3.1 The Santos Operational Forecasting System

The numerical model used for this work is the SOFS, a forecasting system based on the Princeton Ocean Model[1] named POM-rain module that provides forecasts of currents, SSH, salinity and temperature up to three days ahead for the Santos-Sao Vicente-Bertioga Estuarine System.

The SOFS model is based on the Navier-Stokes equation, considering Boussinesq approximation, hydrostatic pressure and an incompressible fluid. The system deals with the following equations:

$$\begin{cases} -\frac{\delta\eta}{\delta t} = \int_0^H \frac{\delta u}{\delta x}\delta z + \int_0^H \frac{\delta v}{\delta y}\delta z, \\ \frac{\delta u}{\delta t} + \vec{V}\cdot\nabla u - fv = -\frac{1}{\rho_0}\frac{\delta p}{\delta x} + \frac{1}{\rho_0}\nabla\tau_x, \\ \frac{\delta v}{\delta t} + \vec{V}\cdot\nabla v + fu = -\frac{1}{\rho_0}\frac{\delta p}{\delta y} + \frac{1}{\rho_0}\nabla\tau_y, \\ \frac{\delta w}{\delta t} + \vec{V}\cdot\nabla w = -\frac{1}{\rho_0}\frac{\delta p}{\delta y} + \frac{1}{\rho_0}\nabla\tau_y - \frac{\rho}{\rho_0 g}. \end{cases} \quad (1)$$

where $\eta$ is the SSH and $\vec{V} = [u, v, w]$ are water velocities in a Cartesian coordinate system where $x$ and $y$ axis are horizontal, and the $z$ axis is vertical. Other parameters of this equation are the total water column depth $H$, Coriolis acceleration $f$, water density $\rho$, water reference density $\rho_0$, gravitational acceleration $g$, pressure $p$ and stresses $\tau_i = [\tau_{ix}, \tau_{iy}, \tau_{iz}]$ in the direction $i$ due to both shear (such as viscosity and wind stress) and Reynolds stress.

The first equation is the continuity equation, assuming water incompressibility; it indicates that the variation in water elevation at a given point is equal to the difference between the volume of water that enters and exits the water column at that point. The three remaining equations are the Navier-Stokes momentum balance in three directions, considering advection and Coriolis accelerations in the left side of the equation and pressure gradient, stresses and buoyancy forces in the $z$ direction in the right side.

SOFS works with two grids, one larger encompassing the coastal region from Southeast Brazil, and a nested grid

[1] http://www.ccpo.odu.edu/POMWEB/

of finer resolution, encompassing the Santos-São Vicente-Bertioga Estuarine System, as shown in Figure 3.

For the coarser grid, this model incorporates atmospheric boundary conditions from the Center for Weather Forecasts and Climate Studies (CPTEC, Portuguese acronym), currents in the open boundary is obtained from the Copernicus Marine Environment Monitoring Service Mercator (CMEMS), and tides in the boundary were obtained by astronomical components for the region. Boundary conditions for the nested grid are obtained from the coarser grid. It uses a three-dimensional grid with Sigma vertical coordinates and Arakawa C-grid for horizontal coordinates.

The SOFS model is also split into two modes, an external mode where the 2D equations are solved considering mean values for the entire water column and that is solved using a faster time-step of 0.8s, and an internal mode where the 3D equation system is solved in a time-step of 4 seconds.

The historical data available for this system is composed of daily forecast events, each one consisting of 8 forecasting steps with a time-step of 3 hours. The initial forecast step is around midnight GMT and the last one is around 21:00 GMT. The forecast is available for all grid points of the system, as shown in Figure 3, but for this work we selected only the points closest to the measuring stations as shown in Figure 1. The selected period is between January $1^{st}$, 2019 and March $19^{th}$, 2021.

### 3.2 In Situ Measured Data

The measured current velocities and SSH used in this project were acquired from Sontek SL Acoustic Doppler Current Profilers (ADCP) installed in the channel; other measurements are obtained from a weather station installed in station P4 (see Figure 1). The measurements available are 5 minute averages of surface current speed and direction for all the 5 stations, 10 minutes averages of SSH for stations P2 through P5, and 5 minutes average, minimum and maximum values of wind, temperature, rainfall and relative humidity as well as average atmospheric pressure for the P4 station.

Data collection in the channel is carried out by the Santos Port Authority. Over time, sensors have been installed in the region to increase spatial coverage and the variety of data available. For this work, we selected the same time window used for SOFS, covering between January $1^{st}$, 2019 and March $19^{th}$, 2021, except for SSH at stations P2, P3, P5, where the measurement started December $1^{st}$, 2019, and for SSH at station P1 where no measurement is available. Data coverage for the selected time range depends on the data type and station. For example, the weather station at P4 has about 4.8% missing data for wind, temperature, and rain, and 1.5% for SSH in the selected time window.

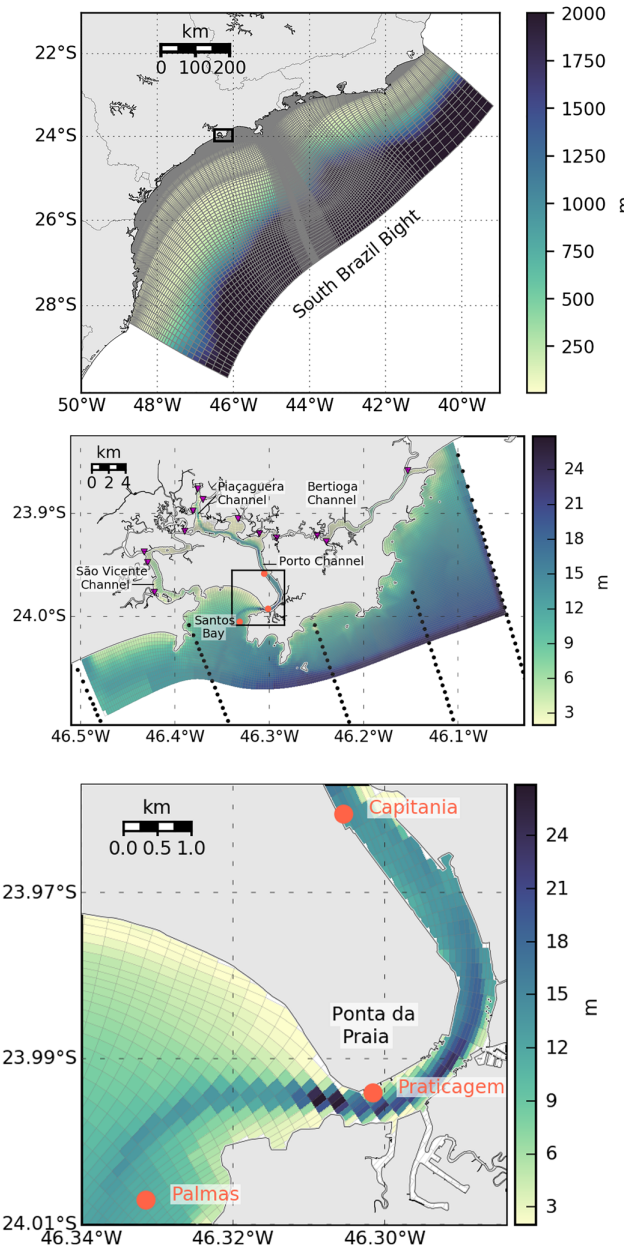


Figure 3: Top: Coarser grid of the SOFS encompassing the South Brazilian Bight with indication of the localization of the nested grid. Middle: nested fine grid for the Santos-Sao Vicente-Bertioga Estuarine system. Bottom: detail of the grid in the entrance of the Santos Channel, the points shown as Palmas, Praticagem and Capitania are respectively stations P5, P4 and P3 in Figure 1. Source: [Costa *et al.*, 2020].

### 3.3 PIML training phase

In the training phase of each ML model, following the supervised paradigm, a training dataset is used consisting of pairs $\langle input\ variables\ ,\ desired\ value\ of\ target\ variable \rangle$.

The measured target variable — either current speed, current direction or SSH — for station $Pn$, with $n \in \{1, 2, 3, 4, 5\}$, in time step $t$, with $t \in \{0, 1, 2, ..., 7\}$, is given by $target_{P_n,t}$. The ML model uses as input variables for the training phase the weather data observed at station P4 at the beginning of the forecast, and the SOFS forecast error for the respective target variable, station, and respective forecast step, $SOFSerror_{target,P_n,t}$. The SOFS forecast error for the respective target variable, station, and respective forecast step, $SOFSerror_{target,P_n,t}$ is given by the difference between the value predicted by SOFS for the target variable,

| Measured Data | | | SOFS Forecast Data | | | | | | | | | Target Variables |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wind Spd. (m/s) | ... | Precip-tation (mm) | Forecast step (One-Hot encoded) 0 | 1 | ... | 7 | Curr. Spd. (m/s) | Curr. Dir. (°) | SSH (m) | SSH temporal difference (m) | SSH spatial difference (m) | Current speed error (m/s) |
| 3.67 | ... | 0 | 1 | 0 | ... | 0 | 0.35 | 92.3 | 2.11 | 0.40 | 0.06 | -0.03 |
| 3.67 | ... | 0 | 0 | 1 | ... | 0 | 0.24 | 85.6 | 1.71 | 0.56 | 0.07 | -0.06 |
| 3.67 | ... | 0 | 0 | 0 | ... | 0 | 0.13 | 248.6 | 1.15 | 0.12 | 0.24 | 0.08 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

Table 1: Example of the training dataset structure for current speed as the target variable. Some measured data columns were omitted for brevity.

$SOFS_{target,P_n,t}$ and the measured value of the same target variable at the respective station, $target_{P_n,t}$. For example, for the SSH target variable and station P5, comes:

$$SOFSerror_{SSH,P_5,t} = SOFS_{SSH,P_5,t} - SSH_{P_5,t}.$$

Once the training database is built, given by pairs of the type

$$\langle \{weather - data_{t=0}, SOFS - data_{P_n,t}, target_{P_n,t}\}, SOFSerror_{target,P_n,t} \rangle,$$

the ML model is trained until a stopping condition is met.

The ML model trained from the PIML model can then be used in the operation phase.

### 3.4 PIML operation phase

In the operating phase, the ML model receives as input the same variables used as input in the training phase, and provides as output the SOFS error estimate for the respective target variable, measurement station and prediction time step, $SOFSerror_{target,P_n,t}$.

However, now the ML model output is subtracted from the prediction made by SOFS in the same step, for the same station and target variable, $SOFS_{target,P_n,t}$, resulting in a corrected prediction for the target variable as the output of the PIML system, $target_{P_n,t}$:

$$target_{P_n,t} = SOFS_{target,P_n,t} - SOFSerror_{target,P_n,t}.$$

## 4 Experimental Setup

By merging both the SOFS and measured data it is possible to assemble a single training dataset for each combination of station $P_n$ and target variable to train the ML model. An example of the dataset structure is shown in Table 1. Each row of the training dataset contains one forecasting step of the SOFS model for SSH and currents, the difference in forecasted SSH between the current and next SOFS step, the difference in forecasted SSH between the current and previous station in the channel, the forecast step in One-Hot Encoding, the measured values of wind, temperature, pressure, precipitation, relative humidity, and SSH in station P4 for the initial forecast step of SOFS, and the target variables. We decided to add the temporal and spatial SSH differences due to the SSH importance in driving the currents in the channel.

To obtain the target variables, as explained in Section 3.3, we subtract the measured values from the values predicted by the SOFS model, taking into account the closest measurement taken in time to the respective SOFS forecast step, within a maximum acceptable difference of 30 minutes. If there was no measurement available within 30 minutes of the forecast step time, we discarded the respective full day. Other data treatments would be possible, but our choice led to the desired analysis.

The direction error is expressed in the range between -180° and 180°, taking the angle wrapping around 0°(0°=360°).

The measured data in each row is the latest measurement available right before the first forecast step of the daily forecast event. If there is no measurement up to 30 minutes before the first forecast step, we discard the entire day. The measurement data that is used as input is composed of all variables measured in the weather station and the SSH variable at the station P4.

The ML model selected to predict the SOFS error is a Random Forest Regressor (RF), available in the Python Package Sklearn, due to the simplicity of tuning its hyperparameters and its characteristic of averaging the target variable values seen in the training dataset, avoiding predicted errors above what has been seen in the past. In a previous work [Moreno *et al.*, 2022], we used Quantile Regression Forests. However, we fond in that previous work that the obtained uncertainty range was too wide to be useful, so we have decided to move to the RF model.

We ran a 5-fold cross validation using random search to find the best hyperparameters of the RF model. The random search approach was performed with 50 random samples at the intervals shown in the table 3. We used the inital 80% of the dataset for the cross-validation and model training, and reserved the remaining 20% for testing.

The hyperparameter optimization was carried on an Intel i7-11800h, taking on average 50 minutes to cross-validate the model for 50 points in the hyperparameter space.

The random search selected the hyperparameters that maximized the Index of Agreement (IOA). The IOA is given by

$$IOA = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(|P_i - \overline{O}| + |O_i - \overline{O}|)^2}, \qquad (2)$$

for a sequence of $n$ observations $O_i \in \{O_1, O_2, ..., O_n\}$ with mean $\overline{O}$, and predicted values $P_i \in \{P_1, P_2, ..., P_n\}$.

| Station | SOFS IOA | | | SOFS+RF IOA | | |
|---|---|---|---|---|---|---|
| | Current | | | Current | | |
| | Speed | Direction | SSH | Speed | Direction | SSH |
| P1 | 0.409 ±0.151 | 0.334 ±0.175 | - | 0.370 ±0.133 | 0.581 ±0.170 | - |
| P2 | 0.528 ±0.193 | 0.636 ±0.281 | 0.922 ±0.113 | 0.516 ±0.195 | 0.662 ±0.288 | 0.933 ±0.115 |
| P3 | 0.563 ±0.203 | 0.688 ±0.159 | 0.880 ±0.141 | 0.556 ±0.200 | 0.671 ±0.162 | 0.886 ±0.137 |
| P4 | 0.591 ±0.231 | 0.754 ±0.185 | 0.945 ±0.051 | 0.622 ±0.227 | 0.783 ±0.181 | 0.958 ±0.047 |
| P5 | 0.483 ±0.193 | 0.585 ±0.221 | 0.944 ±0.083 | 0.553 ±0.182 | 0.668 ±0.206 | 0.953 ±0.076 |

Table 2: Index of Agreement obtained with the combination of SOFS and Random Forests, compared with results obtained with the SOFS model alone.

| Hyperparameter | Interval | Steps |
|---|---|---|
| Number of features | 3 - 15 | 2 |
| Max. tree depth | 5 - 60 | 5 |
| Min. samples for split | 2 - 42 | 4 |
| Min. samples in leaf | 1 - 25 | 4 |
| Number of trees | 50 - 1500 | 50 |
| Loss Metric | MAE, MSE | - |

Table 3: Hyperparameters types and ranges tuned for the post processor. Random search was made inside the interval shown in this table. MAE and MSE stand for "Mean Absolute Error" and "Mean Squared Error", respectively.

The training dataset was then used to train the RF model with the best hyperparameters found in the Cross-Validation. The model was then evaluated in the test dataset, in the operation phase desribed in Section 3.4. The results os the evaluations carried out are described in the following.

## 5 Results

We used the Wilcoxon test to decide whether the distribution of values forecasted by the SOFS model and the PIML model that combines SOFS with the RF model (SOFS+RF) are different.

The Wilcoxon signed-rank test is a non-parametric hypothesis test that is used to verify if two distributions are different in a statistically significant way. To measure the model performance we decided to use the Index of Agreement (IOA), a typical performance metric used to evaluate metocean forecasting models. The IOA metric was calculated for each forecast day in the test dataset. IOA average and standard deviation is shown in Table 4.

The results show that there is an increase in the IOA when the PIML model is employed. The amount of the increase depends on the parameter of interest and the location of the station. For some stations, the improvement in the IOA for the current direction was marginal or non-existent. The improvement obtained for the SSH variable was also small, but this may be due to the high IOA values obtained using the SOFS model, which already shows very good effectiveness, leaving only a small margin for improvement. Current speed prediction using the PIML system, on the other hand, showed a more consistent increase in the IOA metric for all stations.

We also have calculated the Root-Mean Square Error (RMSE) for both the SOFS model and the PIML system.

| Station | SOFS | | | SOFS+RF | | |
|---|---|---|---|---|---|---|
| | Current | | | Current | | |
| | Speed (m/s) | Dir. (°) | SSH (m) | Speed (m/s) | Dir (°) | SSH (m) |
| P1 | 0.089 | 117.9 | - | 0.071 | 91.5 | - |
| P2 | 0.126 | 82.1 | 0.137 | 0.109 | 75.6 | 0.116 |
| P3 | 0.187 | 79.3 | 0.189 | 0.166 | 65.4 | 0.196 |
| P4 | 0.178 | 72.3 | 0.125 | 0.159 | 62.4 | 0.105 |
| P5 | 0.090 | 86.8 | 0.125 | 0.074 | 75.7 | 0.109 |

Table 4: Comparison of Root-Mean Square Error for both SOFS and SOFS+RF.

RMSE decreases for all variables and stations when using the PIML model, except for SSH predictions in the P3 station. One possible explanation for this behavior is a distribution shift between the training and test data.

We also have calculated the Wilcoxon test to verify whether the distribution of predictions obtained with SOFS+RF is different from the SOFS model. If we consider the typical threshold of $p \leq 0.05$ to reject the null hypothesis, we can observe that in the majority of cases the distribution obtained with the SOFS+RF system is different from the one produced with the SOFS model. It can be claimed that the PIML system (SOFS+RF) produces statistically better results than the SOFS model.

| Station | Wilcoxon Test | | |
|---|---|---|---|
| | Current | | |
| | Speed | Direction | SSH |
| P1 | $< 0.001$ | 0.491 | - |
| P2 | $< 0.001$ | $< 0.001$ | $< 0.001$ |
| P3 | 0.001 | $< 0.001$ | $< 0.001$ |
| P4 | $< 0.001$ | $< 0.001$ | $< 0.001$ |
| P5 | $< 0.001$ | 0.055 | $< 0.001$ |

Table 5: P-values for the Wilcoxon test for the distribution of predictions of SOFS and SOFS+RF.

We have also employed histograms to visualize the distribution of forecasts errors of the SOFS and the SOFS+RF, as shown in Figures from 4 to 8. As can be seen, the use of the RF post-processor changes the error distribution. Ideally the post-processor would generate a narrow distribution centered around zero. The error distributions obtained by application

| Station | SOFS | | | SOFS+RF | | |
|---|---|---|---|---|---|---|
| | Current | | | Current | | |
| | Speed (m/s) | Dir. (°) | SSH (m) | Speed (m/s) | Dir (°) | SSH (m) |
| P1 | 0.037 | 19.2 | - | 0.001 | 16.4 | - |
| P2 | 0.030 | 22.7 | 0.010 | 0.022 | 18.9 | 0.005 |
| P3 | 0.050 | 23.6 | 0.050 | 0.042 | 10.6 | 0.097 |
| P4 | 0.054 | 18.1 | 0.034 | 0.041 | 1.5 | 0.019 |
| P5 | 0.027 | 5.3 | 0.029 | 0.006 | 3.4 | 0.004 |

Table 6: Comparison of the error distribution absolute bias for the SOFS and the SOFS+RF.

of the RF post-processor have on general a mean closer to zero in most cases, or lower bias, specially for SSH forecasts when compared to the SOFS alone. A comparison between absolute biases with and without the post-processing for each station is shown in Table 6, where the bias is computed by calculating the average error obtained in the respective case. The only case where the bias increased by applying the post-processor is for SSH in station P3, the same one that saw an increase in RMSE after application of the post-processor.

The improvement due to the post-processor is most evident in the histograms in cases where the SOFS model has does not produce accurate forecasts, such as the case for current directions in P1 and P5, where the lower current speeds make the direction forecast more prone to error, or in P4, where the SOFS model has a positive or negative error depending if the current is entering or exiting the channel, producing a two peaked distribution. When compared to previous results in the region, the decrease in bias obtained from this post processor is smaller, but this might be due to the fact that the train and test datasets were split into two continuous time-series in this paper, while they were split at random in the previous study.

## 6 Conclusions

In this paper we have presented a Physics-Informed Machine Learning approach to reduce the forecasting error of surface currents and SSH in the Santos Channel, Brazil, by predicting the forecasting *error* of the SOFS model, a numerical model already in use for the region.

In our PIML approach we used Random Forests (RF) to predict the SOFS error by taking into account, as inputs, both the SOFS forecasts and sensor data. An RF model was trained for each combination of target variable (current speed, current direction and SSH) and channel location (points P1 to P5 in Figure 1), resulting in 15 different models. The best hyperparameters for each RF were found by 5-fold cross-validation in a training dataset containing 80% of the data, using the Index of Agreement (IOA) as a metric for performance. Once the RFs are trained, they can be used as a post-processor to correct the SOFS forecasts.

This post-processor was tested on a dataset not previously seen in training, and its performance was measured using IOA and RMSE as metrics. The results show that the use of that post-processor increased IOA in all stations for SSH, in 4 out of 5 for current direction, and in 2 out of 5 stations for current speed, while the RMSE decreased for all combinations of target variables and stations, except in the case of the SSH for the P3 station. One possibility is that the increase in RMSE for that specific case is due to distribution shift, but further analysis is still required to find the root causes.

Several opportunities arise as this investigation continues. As an immediate next step, we intend to improve the proposed PIML architecture with tests and comparisons using other ML models as post-processors. We also intend to experiment with architectural modifications so that more information can be used as inputs to the post-processor, as in this version our post-processor uses measured data obtained only immediately before the first prediction step. Another proposal is to use Long Short-Term Memory (LSTM) Neural Networks or Transformers to encode an arbitrary long time series of measurements taken just before the prediction event into a fixed-length input that will provide more information for predictions.

## Acknowledgements

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant number 2019/07665-4) and by the IBM Corporation. This work is also supported in part by FAPESP (grant number 2020/16746-5), the Brazilian National Council for Scientific and Technological Development (CNPq grant numbers 310085/2020-9, 310127/2020-3, and 312180/2018-7), Coordination for the Improvement of Higher Education Personnel (CAPES, Finance Code 001), and by *Itaú Unibanco S.A.* through the *Programa de Bolsas Itaú* (PBI) program of the *Centro de Ciência de Dados* ($C^2D$) of *Escola Politécnica* of USP.

## References

[Bonavita and Laloyaux, 2020] Massimo Bonavita and Patrick Laloyaux. Machine Learning for Model Error Inference and Correction. *Journal of Advances in Modeling Earth Systems*, 12(12), 12 2020.

[Chapman *et al.*, 2019] W. E. Chapman, A. C. Subramanian, L. Delle Monache, S. P. Xie, and F. M. Ralph. Improving Atmospheric River Forecasts With Machine Learning. *Geophysical Research Letters*, 46(17-18):10627–10635, 9 2019.

[Cho *et al.*, 2020] Dongjin Cho, Cheolhee Yoo, Jungho Im, and Dong Hyun Cha. Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas. *Earth and Space Science*, 7(4), 4 2020.

[Costa *et al.*, 2020] Carine G.R. Costa, José Roberto B. Leite, Belmiro M. Castro, Alan F. Blumberg, Nickitas Georgas, Marcelo Dottori, and Antoni Jordi. An operational forecasting system for physical processes in the Santos-Sao Vicente-Bertioga Estuarine System, Southeast Brazil. *Ocean Dynamics*, 70(2):257–271, 2 2020.

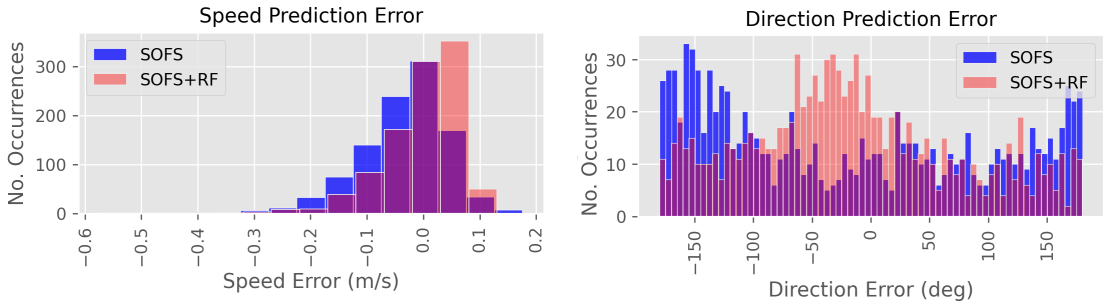


Figure 4: Forecast error distributions for point P1.

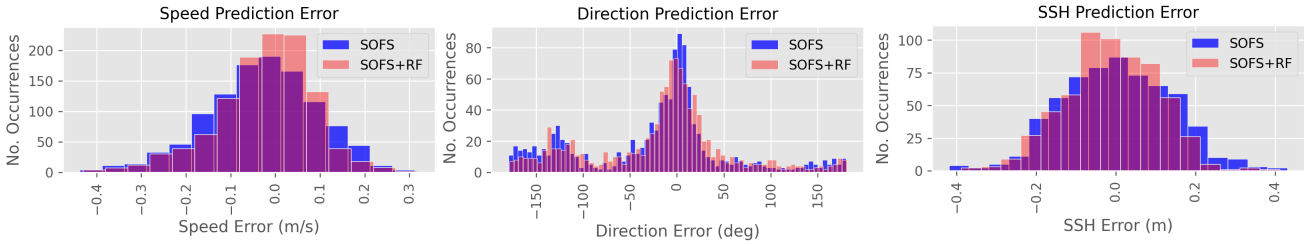


Figure 5: Forecast error distributions for point P2.

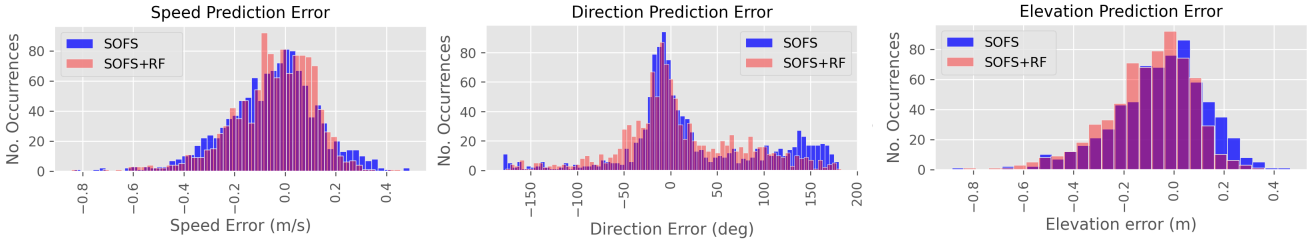


Figure 6: Forecast error distributions for point P3.

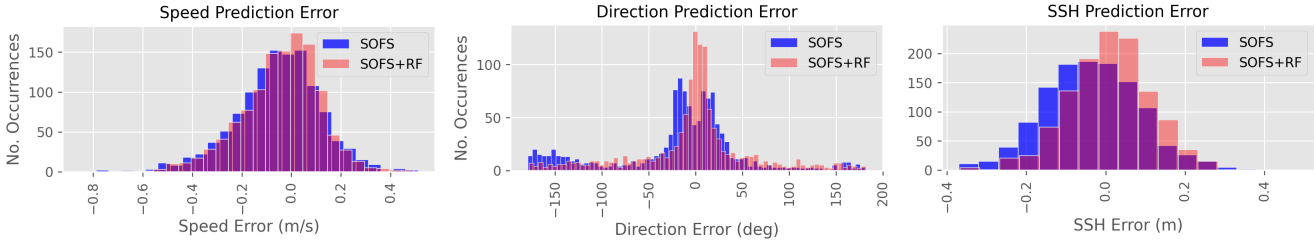


Figure 7: Forecast error distributions for point P4.

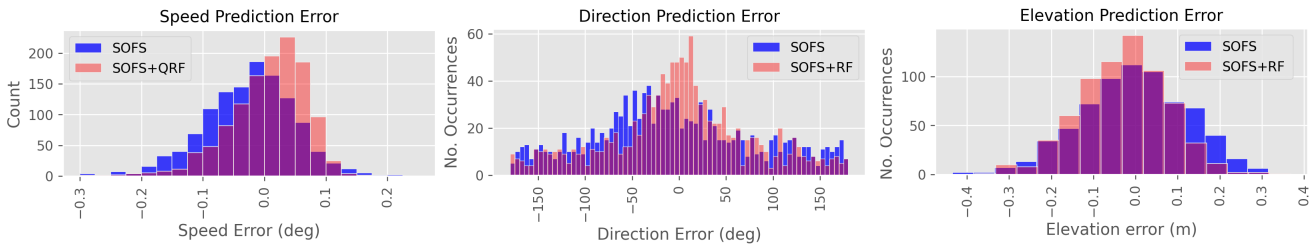


Figure 8: Forecast error distributions for point P5.

[Eccel *et al.*, 2007] E Eccel, L Ghielmi, P Granitto, R Barbiero, F Grazzini, and D Cesari. Prediction of minimum temperatures in an alpine region by linear and non-linear post-processing of meteorological models. *Nonlin. Processes Geophys*, 14:211–222, 2007.

[Isaksson, 2018] Robin Isaksson. Reduction of Temperature Forecast Errors with Deep Neural Networks, 2018.

[Kashinath *et al.*, 2021] K. Kashinath, M. Mustafa, A. Albert, J. L. Wu, C. Jiang, S. Esmaeilzadeh, K. Azizzadenesheli, R. Wang, A. Chattopadhyay, A. Singh, A. Manepalli, D. Chirila, R. Yu, R. Walters, B. White, H. Xiao, H. A. Tchelepi, P. Marcus, A. Anandkumar, P. Hassanzadeh, and Prabhat. Physics-informed machine learning: Case studies for weather and climate modelling, 4 2021.

[Moreno *et al.*, 2022] Felipe M. Moreno, Luiz A. Schiaveto Neto, Fabio G. Cozman, Marcelo Dottori, and Eduardo A. Tannuri. Enhancing the forecast of ocean physical variables through physics informed machine learning in the santos estuary, brazil. In *OCEANS 2022 - Chennai*, pages 1–7, 2022.

[Scher and Messori, 2018] Sebastian Scher and Gabriele Messori. Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, 144(717):2830–2841, 10 2018.

[Vashani *et al.*, 2010] S Vashani, M Azadi, and S Hajjam. Comparative Evaluation of Different Post Processing Methods for Numerical Prediction of Temperature Forecasts over Iran. *Research Journal of Environmental Sciences*, 4(3):305–316, 2010.

[Willard *et al.*, 2020] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating scientific knowledge with machine learning for engineering and environmental systems. *arXiv preprint arXiv:2003.04919*, 2020.

[Xu and Valocchi, 2015] Tianfang Xu and Albert J. Valocchi. Data-driven methods to improve baseflow prediction of a regional groundwater model. *Computers and Geosciences*, 85:124–136, 12 2015.