Journal Name

Crossmark

RECEIVED
dd Month yyyy

REVISED
dd Month yyyy

ARTICLE

# Forecasting Coastal Water Level Extremes from Past Observations with Explainable AI

Avery Wood[1,*], Maike Sonnewald[1] and John L. Largier[2]

[1]Department of Computer Science, University of California Davis, Davis, CA
[2]Coastal and Marine Sciences Institute, University of California Davis, Bodega Bay, CA
*Author to whom any correspondence should be addressed.

**E-mail:** awood@ucdavis.edu

## Abstract

Coastal water level variance exhibits an abundance of normal tidal behavior but occasional meteorological events can result in extreme high or low water levels. Predicting when (and why) these significant changes occur is difficult and requires both physical understanding and insight to modeling techniques that can accurately predict extreme values without sacrificing too much performance during non-extreme conditions. We specifically look at predicting the water level in the lower Petaluma River near its confluence wit San Francisco Bay (California). Classical regression models perform well at modeling normal meteorological effects but fall short during extreme events, which are responsible for flooding of low-lying lands and infrastructure adjacent to the river. We test a hierarchy of machine learning forecasting models suitable for time-series data and show much improved predictability of extreme water level events in the Petaluma River. We test MLP, LSTM, and Transformer models and ensembles of each of the models to further improve consistency and reliability. Finally, to ensure that our models are trustworthy, we measure explainability of input features using Shapley Values (SHAP) and verify feature importance with expert knowledge of the physical processes in this coastal waterway. We show that LSTM can reliably forecast non-tidal effects and thus also water level extremes with errors less than 0.1 m during rare but severe meteorological events.

## 1 Introduction

In both 2017 and 2019, California State Highway 37 was closed for days due to flooding associated with high water levels along the northern shore of San Francisco Bay. Widespread flooding occurred landward of the levees along the lower Petaluma River and Novato Creek. This flood and the consequent disruption of this major highway had a significant economic and environmental impact. Coastal flood events such as this are becoming increasingly common due to the rise in sea level that has been observed over the past several decades (Griggs Report). A critical challenge is to forecast water levels well enough to mitigate the risks associated with coastal flooding. Most work to date has been on numerical simulations and generic scenario projections (e.g., Barnard CoSMoS), rather than actionable forecasts of water level at specific places and times.

Here we use high-frequency water level observations at a specifically vulnerable site on the Petaluma River to forecast water levels on the same time horizon as weather forecasts. This case study is the first step in developing a robust, water-level forecasting model for sheltered coastlines in San Francisco Bay and elsewhere. We previously used a regression model with reasonable success when combined with a hindcast correction to account for low-frequency seasonal and inter-annual effects (Largier et al 2023). Following prior work using Machine Learning (ML) methods to improve forecasting of extreme events (Thirumalaiah et. al.[1], Campolo et. al.[2]), here we test a hierarchy of ML models, which show significant improvements in forecasting water level extremes.

Beyond our interest in the skill of ML forecasting, we are interested in whether this skill is rooted in physical relations, and thus more "trustworthy". Different ML models are based on different strategies (different internal mechanisms) that are suitable for different types of problem. Thus, we test a variety of model types with increasing complexity: Multi-Layer Perceptron (MLP; citation), Long-Short Term Memory (LSTM) (Hochreiter


Figure 1: Left panel: flooding of Highway 37 in 2019 ([3]). Right panel: Map showing Highway 37 (red ellipse) traversing the wetlands on the northern shore of San Pablo Bay, San Francisco Bay. Petaluma River is visible in the top left of the map ([4]).

et. al.[5]), and Transformers (Lopez et. al.[6]). In regard to water level prediction, MLPs have been shown to have longer, more accurate forecasting windows than traditional counterparts such as regression models (Xu et. al.[7]). Similarly, LSTMs have been shown to have superior performance on most time-series data, including hydrological data, when compared with other ML methods (Vizi et. al.[8]). Given the LSTM suitability for our type of data, we expect this model to perform well. Lastly, Transformers have been shown to perform competitively for hydrological forecasts under the right circumstances (Lopez et. al.[6]). Given the newness of transformers, it also seems fitting to include them to measure their performance in real-world forecasting. Each of these models is a step up in complexity from our prior work using a regression model and can be expected to handle nonlinearities better, leading to significantly more robust forecasting of extreme events. Knowing the source of predictive skill is important for gaining trust in a model, which is necessary when used to inform decisions with societal impact such as closure of Highway 37 prior to a anticipated flood event. ML models are "data-driven", like traditional correlation models, in the sense that they take a given dataset and find relations that can be used to extrapolate future conditions. ML models are different to numerical simulations models in that they do not explicitly model the physical drivers. Skepticism around deploying ML models is derived from the "black box" approach in which one does not know the implicit rationale and associated physical mechanisms behind the predictions. Here, we use explainable AI (XAI) to identify which input features are important in the prediction. Specifically, we use Shapley Additive (SHAP) explanations ([9]) to verify the source of skill for predicting anomalies. If SHAP values correspond with what we expect in the physical system, then we gain confidence in the prediction. Confidence also comes from only including input features that are known to be physically important (e.g., onshore winds, low atmospheric pressure, and high river flows are known to elevate water levels). In addition to testing different models we also test them as both single models and as ensembles. The ensembles we specifically create consist of a single model type that is trained multiple times. The forecasts of the individual ensemble members are then aggregated into a single, unified forecast. Lee et. al.[10] and Leutbecher et. al.[11] both demonstrated that ensembles of this nature provide significant improvements to the inherent uncertainty found within most neural network based architectures and greatly improve the understanding of the systems being modeled. While we don't explicitly explore it here, multi-model ensembles have also been shown to enhance the forecasting ability of natural systems. Hagedorn et. al.[12], and the European Center for Medium-Range Weather Forecasts[13] by extension, specifically investigate these and identify that these multi-model ensembles are similarly as capable as their single-model counterparts while also increasing robustness to model overfitting.

Ensembles also exist as a regularization technique. Moradi et. al.[14], and much earlier Bireman et. al.[15], demonstrated that ensembling acts as regularization through the inherent stochasticity associated with training Machine Learning techniques. For our case, we are also interested in the effects of regularization on the forecasting of anomalies/extreme values. Abati et. al.[16] and Wang et. al.[17] both demonstrated that ensembles can effectively improve the predictability of anomalies through the regularization they provide. Essentially, by regularizing the latent space of the models through the ensemble the resultant networks are more capable of identifying separable features between normal and abnormal samples. This ability of the model ensembles is what we are specifically trying to replicate with this work as it enables us to specifically target the extreme values/anomalies in our data and then further explain their causes using XAI techniques like SHAP.

Given that our goal is to specifically forecast anomalies, in this case extreme water level values, the metrics we use and how we validate model performance should be focused on performance at these times. Anomalies are difficult to measure quantitatively as most modern ML models use loss functions to guide their "learning". These loss functions minimize the difference between the model predictions and the actual values, leading to an increase in accuracy as the model is trained longer. However, truly anomalous values (extreme values) lie outside the normal distribution, which represents the typical behavior of the system, and improved prediction of typical behavior may not represent improved prediction of anomalous events. This is true in our case (and for many natural phenomena) with a few anomalous values among an abundance of normal behavior. Clearly a metric like RMSE (a good indicator for overall model fit to data) does not describe the error associated with just the anomalous values. We need to look at anomalous events and describe their error manually (both quantitatively and qualitatively).

In summary, we are interested in addressing the ability of ML models to accurately predict water level extremes (anomalies). Modern ML methods can offer improvements over traditional regressions, specifically when using ensembles of the models. Further, we address the trustworthiness of the tested models by using the XAI method SHAP to examine the importance of each input feature in each of the models. A description of our data and the methods are in section 2. Our results are outlined in section 3 followed by a discussion of the results in section 4. The conclusions are in section 5.

## 2 Methods and data

### *2.1 Data*

We use a 20,000-hour time series of water level elevation observed every 10 minutes from October 2020 to February 2023 (Largier et al 2023). Specifically, we analyze data from a pressure sensor in the lower Petaluma River, one mile from its confluence with San Francisco Bay. Tidal variation in water level is deterministic and well predicted (Pugh 2000) and readily available from tide charts (e.g., NOAA site). Here we remove the predicted tidal signal using the utide package ([18]) in python to focus on non-tidal signal (i.e., "tidal residual") – the deviation of observed water level from predicted water level. This tidal residual signal is expected to explain extreme water levels (anomalies) associated with flooding.

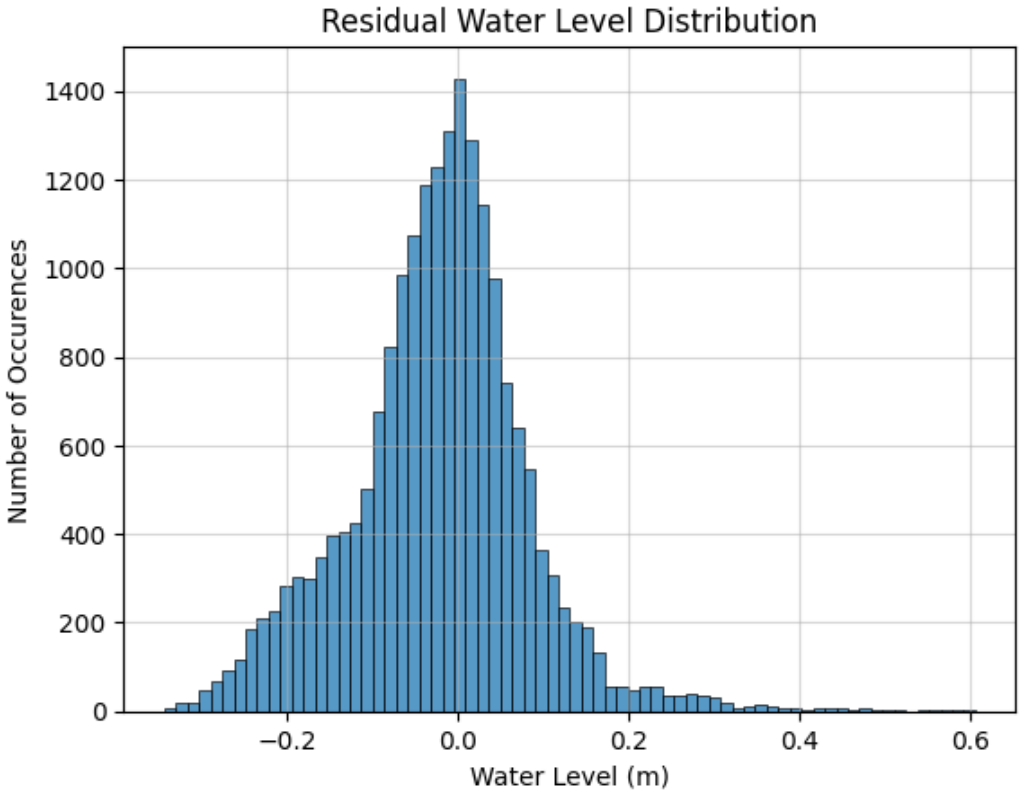



Figure 2: The statistical distribution of residual water level values with mode at 0.0 m and inter-quartile range from -0.1 to +0.1 m. Residual values above +0.18 m are rare, but they may be as large as 0.6 m.

To develop a physically realistic model, we select input features that are known to have a causative relation with water level in bays: barometric pressure, ocean storm surge (represented by ocean winds), local winds, and river inflow rate. Further, to develop a forecast of future water level, we are limited to input features that are themselves reliably predicted by reliable institutions. For example, we use river flow data from the Napa River (USGS site 11458000) as there is no prediction of flow in the Petaluma River; the Napa River is a nearby small watershed that is expected to have highly correlated rainfall and runoff. Ocean storm surge is indexed by coastal winds observed at the offshore NOAA/NDBC buoy 46013, which are well correlated with open-coast water levels in the vicinity of the mouth of San Francisco Bay (Bjorkstedt et al 2015). Barometric pressure data are also from NOAA/NDBC buoy 46026 as atmospheric pressure varies little on spatial scales less than 100km. Local winds over the north shore of San Francisco Bay are well represented by observations at Gnoss Field, an airfield near Petaluma. More information on data properties and sources are given in Largier et al (2023).

### *2.2 Regression*

In prior work we used a simple regression model (Largier et al 2023), which we consider as a baseline. The model has two main components: a multi-input regression model and a hindcast correction. The regression part of the model uses four input features (local wind, ocean wind, barometric pressure, river flow) and the PolynomialFeatures ([19]) function from the sklearn library. This function creates linear combinations of the features up to a specified degree. Largier et al (2023) included quadratic and cubic features, which improved the ability of the model to fit to out-of-distribution points such as the extreme values in our data. However, adding more terms to the model increases the risk of overfitting. The hindcast portion of the regression model adjusts the model for the error in prior predictions (averaged over twelve hours), which corrects for low-frequency background effects such as seasonality and inter-annual variability not represented by the high-frequency meteorological effects that are our primary interest. This hindcast correction is slowly varying (time scales longer than a week).

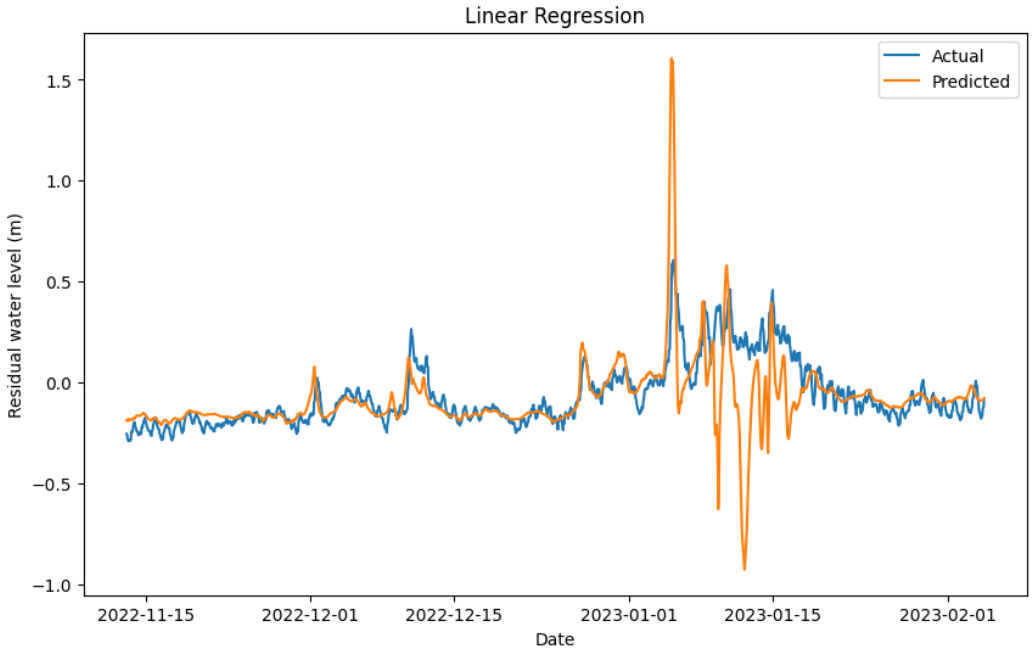



Figure 3: The regression-based prediction (orange line), as reported by Largier et al (2023), and observed residual values (blue line) for target period from 12 November 2022 to 4 February 2023.

### *2.3 MLP*

We chose the Multi-Layer Perceptron (MLP) model as a first step into the world of non-linearity that modern ML methods handle well. Our MLP is a simple two-layer network with an input size of $4x1$ that represents the four features: Napa River flow, barometric pressure, ocean wind, and local wind (Gnoss Field). Our output is the residual water level in the Petaluma River at any given time step.
We used a two-layer MLP with 50 nodes in each layer. We used ReLu activation functions in both of the layers and a linear activation function in the output. We used the Adam optimizer and MSE as our loss function. This is the configuration we found to have the best performance, without making the model unnecessarily complex. Our features were processed using the same PolynomialFeatures function (degree = 3) as the linear regression model and our output had the same rolling hindcast applied to it. We also tested $R^2$ and SSE as loss functions.

### *2.4 LSTM*

The next model we tested was Long Short-Term Memory (LSTM), which are well known for their performance in forecasting tasks, yet only recently have they been applied to water level forecasting. The LSTM we use consists of a LSTM layer with ten nodes followed by a dense layer with ten nodes and an output layer. We used a simple model as increasing the size and amount of LSTM layers greatly increases the complexity and in testing varying sizes of the LSTM and dense layers we found no significant improvement – likewise, we did not see significant improvement when adding additional LSTM layers. Keeping the model simple is important consideration as we use a post-hoc SHAP method to explain our models and their predictions. Shapley values are computationally expensive method of explanation and therefore we benefit by not increasing model complexity. We tested the same loss functions as our MLP including MSE, SSE, and $R^2$. Further, we trained for 20 epochs and until convergence using the Adam optimizer. Lastly, we included the hindcast from the regression model to overcome fitting difficulties that we believe the LSTM can not overcome by itself – discussed later.

### *2.5 Transformer*

The third ML model we tested was a transformer. Transformers are well known for their application in natural language, but they have also seen success in time series forecasting. Namely architectures like Informer (Zhou et. al.[20]) and, more recently, Powerformer (Hegazy et. al.[21]) have shown that Transformers can be modified and applied effectively

in long time-series forecasting. We use a simple transformer architecture that follows the common framework of an encoder with multi-head attention followed by a processing, feed-forward layer and the output layers. We first process the input into a positionally encoded embedding layer to match our input sequences to the size of the transformer. Our positional encoding uses sequences of length 64, corresponding to a 64-hour look-back window. This time span provided the best qualitative results for predicting extreme water level values. Longer look-back windows elongate the predictions while shorter look-back windows sacrifice performance. This encoding is then fed into our encoder which consists of standard multi-head attention as found in the original transformer architecture (Vaswani et. al.[22]). Our processing layer is a single dense layer using ReLu activation functions, and our output layer produces a single prediction value for the water level at the next time stamp. We trained our transformer for 35 epochs, until convergence, using the Adam optimizer and MSE as our loss function.

### *2.6 Ensembles*

For each of the above models we generated ensembles. We stochastically trained each model type 10 times and then aggregated the results. Similarly to the individual models, we trained them until convergence using the Adam optimizer and tested using MSE, SSE, and $R^2$ loss functions. Generating ensembles produces a range in the performance of each model and highlights model performance in regards to extreme values (anomalies) water level residuals. This is useful in determining the efficacy of each model type in predicting anomalies as well as their consistency in their predictions.

### *2.7 SHAP*

For each model we calculated SHapley Additive exPlanations (SHAP) values to quantify the importance of each input feature in the prediction. SHAP values are calculated by removing a feature and seeing how that affects the prediction. Doing this for all features and all predictions is a costly process but necessary to understand and validate the model in terms of physical processes. SHAP values show how important each feature is in contributing to the residual water level. We can then see what accounts for extreme values - is it just one feature, or a combination of features? We can also use SHAP to verify that the ensembles we create are accurately capturing all extreme values. It is possible that each ensemble member captures different extremes to different degrees and together they provide a fuller picture of what is important for each residual water level event.
We calculate the SHAP values using the shap ([9]) library in python, using two explainers: deep explainer and permutation explainer. These methods are approximations of the traditional SHAP values but show superb accuracy when the explainer is trained on a sufficiently large training sample. Deep explainer was our primary explainer. Permutation explainer was used with the LSTM as deep explainer could not be used due to the input size. For the transformers and MLPs we were able to train the SHAP explainer using the full training set. For the LSTM we used a subset of 100 samples for training the permutation explainer. For all of the ensembles we trained each ensemble member and calculated their SHAP values individually. We then aggregated all SHAP values to get the mean value presented in Fig. [all shap].

## 3 Results

The distribution of tidal residual values shows a quasi-normal distribution (Figure 2) and most values are well predicted by all of the methods explored (Table 1). RMS error is between 0.104 and 0.172 m, which is adequate for most operational needs, and MAE is similarly small (0.053 to 0.151 m). While ML models show lower RMSE than the regression model, and the MLP Ensemble performs the best, the differences are small. This consistency in performance is due to performance on normal behavior. However, our interest is in extreme values, which are rare and anomalous (in Figure 2, see tail of distribution with residual values greater than +0.18 m).

| Model | RMSE (m) | MAE (m) |
|---|---|---|
| Regression | 0.172 | 0.0840 |
| Single MLP | 0.132 | 0.0567 |
| MLP Ensemble | 0.104 | 0.0526 |
| Single LSTM | 0.161 | 0.1465 |
| LSTM Ensemble | 0.163 | 0.1505 |
| Single Transformer | 0.155 | 0.1282 |
| Transformer Ensemble | 0.140 | 0.1259 |

Table 1: Root-mean-square error (RMSE) and mean absolute error (MAE) for all models tested.

To avoid waiting decades to accumulate enough data for conventional statistical assessment of anomalous events, we address this issue by reviewing model performance during a series of extreme water levels observed during spring high tides in January 2023. After removing the tidal signal, we see several weather-related events that account for large positive residuals in December 2022 and January 2023 (Figures 3, 4, 5 and 6). Observed tidal residuals are close to -0.2 m between events, and during events the residual is often +0.2 m, increasing briefly to +0.6 m on 5 January (i.e., a range of about 0.8 m). The regression model (Figure 3) performs well in general with discrepancies between prediction and observation that are less than 0.2 m, even during the high-value event mid-December. However, this method struggles to predict the extreme event on 5 January based on correlations developed from prior data - the model-observation discrepancy is about 1.0 m. It also struggles to predict residual values for the following 10 days as input feature values change significantly and rapidly, e.g., a large negative discrepancy of about -1.0 m occurs on 12 January.
The predictions for the MLP ensemble model are shown in fig. 4. This model performs better than the regression model in the test period. In general, the prediction is within 0.1 m of the observed values (similar to the RMSE), including both negative and positive biases. However, the model also struggles with the extreme event on 5 January (positive discrepancy of about 1.0 m), but does better over the subsequent 10 days (discrepancies less than 0.5 m). The standard deviation of the ensemble is negligible most of the time, and remains less than 0.1 m during all events except the 5 January event, when it increases to about 0.5 m, indicating significant model uncertainty at that time.

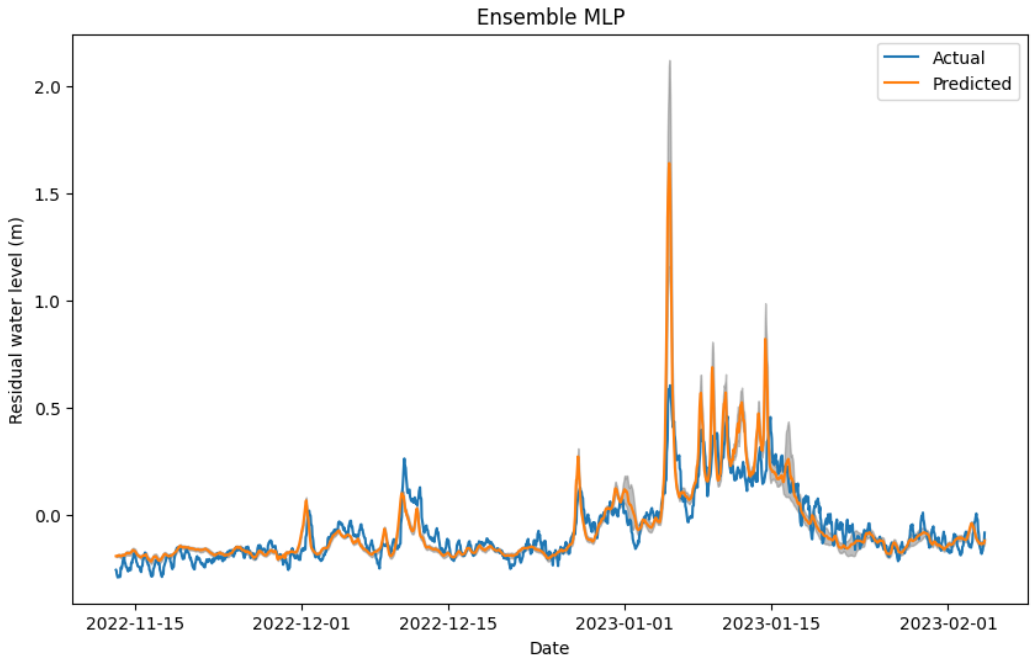



Figure 4: The MLP ensemble prediction (mean of ensemble members; orange line) and observed residual values (blue line). The standard deviation of the ensemble is shown as a gray zone representing one standard deviation around the mean.

The predictions for the LSTM ensemble model are shown in fig. 5. Predictions are

generally within 0.1 m of observed values. However, this model performs much better during the peak in residual values on 5 January with a discrepancy of less than 0.2 m. The LSTM ensemble model under-predicts this peak, but it exhibits a small standard deviation (i.e., high confidence in the model prediction). Indeed, the standard deviation remains low for the entire 2.5-month period and only briefly increases above 0.05 m in late January.

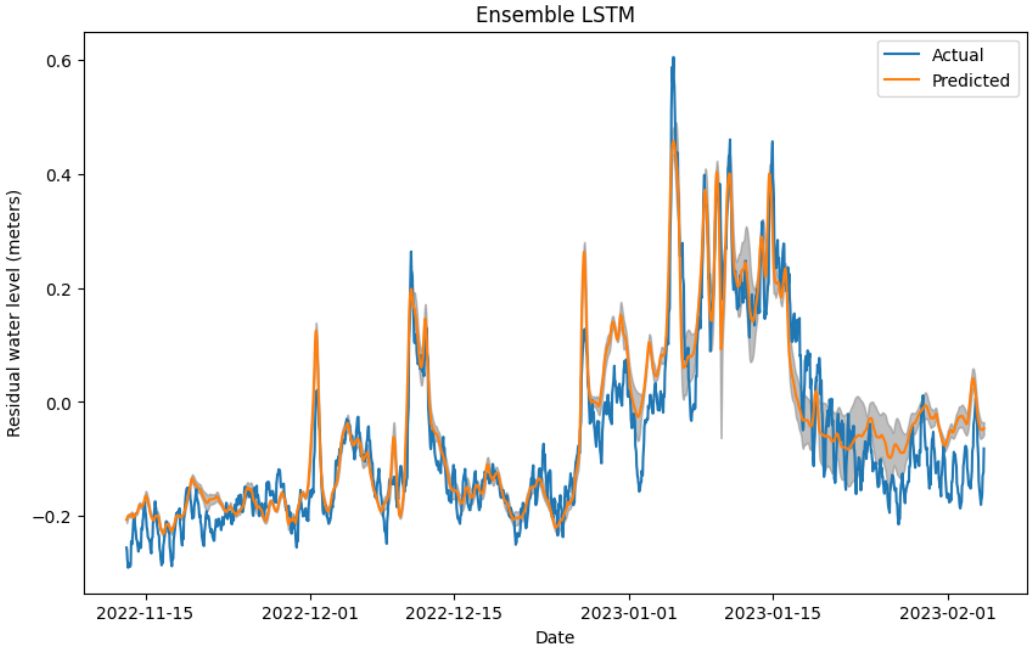



Figure 5: The LSTM ensemble prediction (mean of ensemble members; orange line) and observed residual values (blue line). The standard deviation of the ensemble is shown as a gray zone representing one standard deviation around the mean.

The predictions for the Transformer ensemble model are shown in fig. 6. While the model does reasonably well during high residual values in January (and the mid-December event), with discrepancies less than 0.2 m, it exhibits a persistent positive discrepancy of about 0.2 m between events. Nevertheless, the standard deviation is very small during this period, showing that ensemble member predictions are very similar.

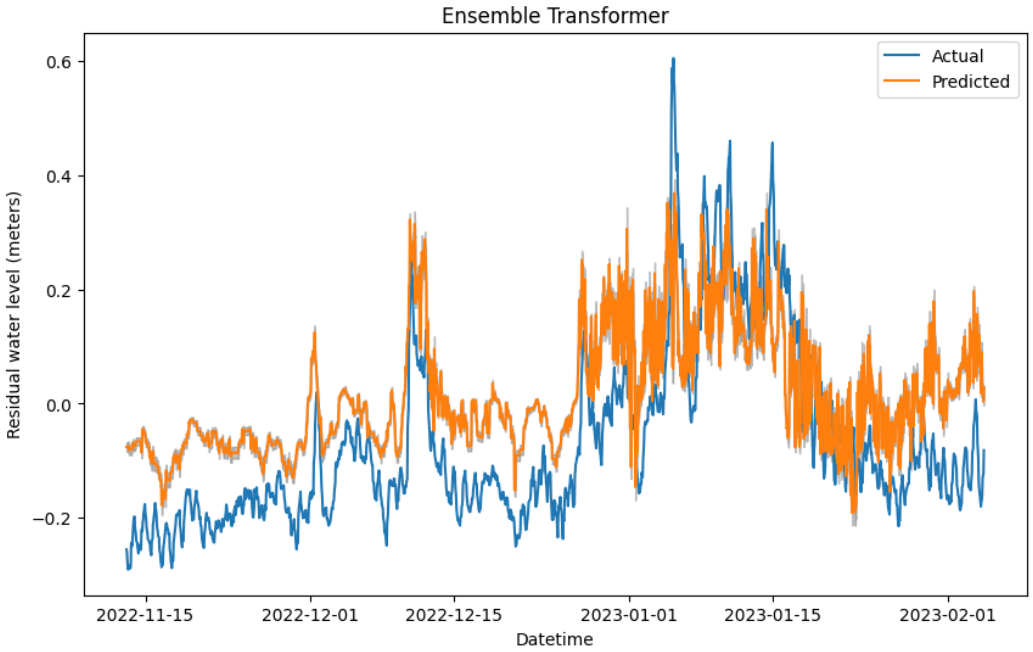
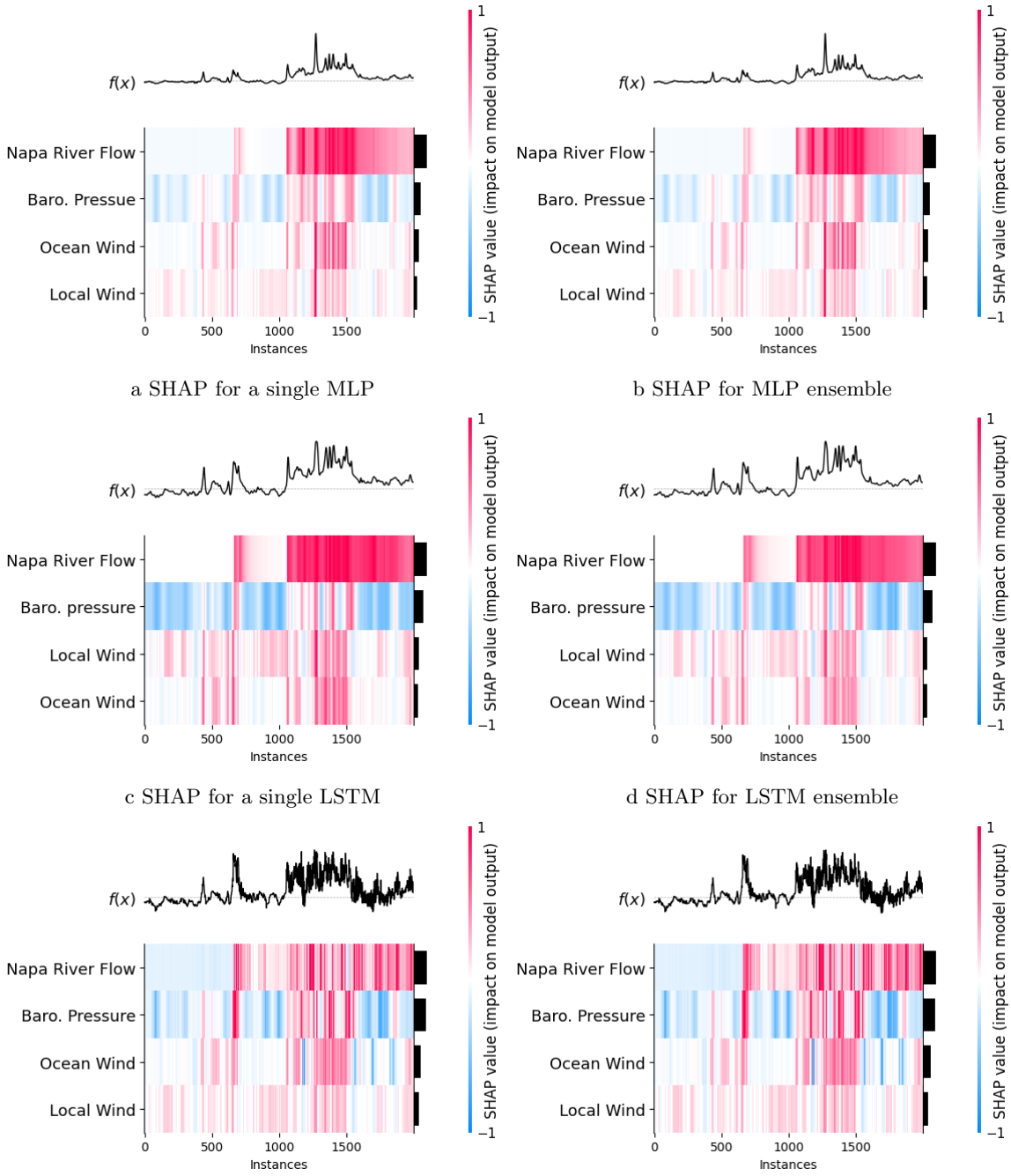



Figure 6: The Transformer ensemble prediction (mean of ensemble members; orange line) and observed residual values (blue line). The standard deviation of the ensemble is shown as a gray zone representing one standard deviation around the mean.

The SHAP values for each of the models and their respective ensembles are shown in Fig. 7. For all models, the Napa River flow and barometric pressure are the most important input features. Local wind and ocean wind are generally less important, but they are significant during specific events (e.g., January 2023, data points 1300-1500). Consistent with physical processes, the influence of Napa River flow is dominantly positive, i.e., higher water levels occur when Napa River and thus also Petaluma River are flowing more strongly. Specifically, the highest residuals align with persistent high SHAP values for Napa River flow. Further, the high SHAP values for river influence are slowly varying, rising fast and falling slowly as do flow rates in rivers. The relation with barometric pressure is predominantly negative, but also consistent with physical processes, as high pressure is associated with lower water levels. These SHAP values show higher frequency variability, representing the higher frequency weather-related variability in atmospheric pressure. Interestingly, pressure SHAP values are not always negative. Likewise, local and ocean wind values can be negative or positive, but primarily they are positive as we chose to define wind inputs as the vector component most correlated with high water levels in the Bay (local wind) or ocean (ocean wind), i.e., onshore winds result in higher water levels either at the shoreline in the Bay (local winds) or the shoreline in the ocean, at the Bay mouth (ocean winds). Again, wind influences exhibit higher frequency variability, representing weather-related variability in winds. On many days it appears that local wind and ocean wind work together (correlated positive values).




a SHAP for a single MLP

b SHAP for MLP ensemble

c SHAP for a single LSTM

d SHAP for LSTM ensemble

e SHAP for a single Transformer

f SHAP for Transformer ensemble

Figure 7: SHAP values for each model, both single runs (left column) and ensembles (right column). MLP values in top row, LSTM results in middle row, and Transformer values in bottom row. In each panel, the model prediction is shown by a black line labeled f(x) and SHAP values are indicated by colors. Data are shown for date to date. Features are ordered by overall importance, denoted by the black bars on right of each panel.

In the results for MLP and LSTM, the alignment of persistent high SHAP values for the Napa River flow with peak water levels is the dominant signal. Barometric pressure is less important in the MLP model than others, and wind influences are always positive. In the LSTM model, the persistent river influence is most striking, but barometric pressure is also important, accounting for short-term fluctuations in residual values. Wind effects are again weaker, but always positive.
The results for the transformer model are different. The Napa River flow SHAP values fluctuate rapidly, associated with high-frequency variability in the transformer predictions that appear to be noise when compared with observed residuals (Figure 6). It is intriguing that the Napa River flow remains important towards the end of the study period, which is

not found in the other models. As in other models, high barometric pressure correlates well with low residual values in the residual while still also being positively correlated with extreme values.
Although not so clear for the transformer model, the concurrence of feature influences is important in explaining the 5 January peak and other peaks in residual values. This makes sense as one would expect extreme values to be the product of multiple features working together in phase.
To visually report the model skill in Fig. 8, we show the bias between the actual residual and the residual forecast by the respective models.

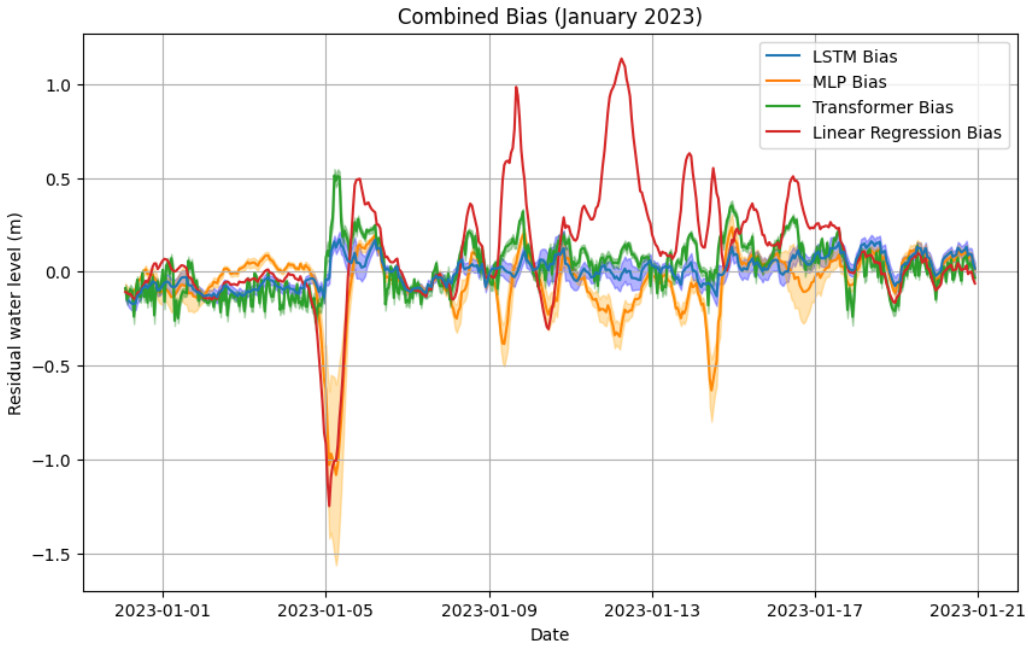



Figure 8: This figure shows the residual water level mean biases within the month of January 2023. The standard deviation across ensemble members is included as well to demonstrate the confidence of each ensemble models predictions throughout the time frame. Interestingly we can see that the MLP shows the highest uncertainty when predicting extreme values such as the one on January 5th. Both the LSTM and Transformer show lower uncertainty throughout all of their predictions.

The bias for the ensemble LSTM is seen as the blue lines in figure 8. The RMSE and MAE for a single run of this model are 0.161 (m) and 0.1465 (m) respectively (Fig. 1). The overall RMSE and MAE for the LSTM ensemble are 0.163 (m) and 0.1505 (m) respectively (Fig. 1). Looking further at the bias we can see that the blue lines in figure 8 lie much closer to zero throughout the predictions compared to the previous MLP ensemble and the existing Linear Regression model. Specifically, in Fig. 8 we can see that in our targeted time frame the performance of the model is quite good.
For a single Transformer our RMSE and MAE are 0.155 (m) and 0.1282 (m) respectively. For the Transformer ensemble these slightly improved to a RMSE of 0.140 (m) and an MAE of 0.1259 (m). The Transformer ensemble's bias is represented by the green lines in figure 8. In these figures we can see that the transformer had a systemic negative bias throughout most of the predictions. Interestingly, in our target time frame the model performs far better with a bias closer to zero (Fig.8).
Notice that the model is substantially noisier than either of the previous two model types. Also, the standard deviation is small throughout the entire time-series indicating that the transformer ensemble is quite certain that this noise is present. We speculate this may be due to a lack of data, but it is equally likely that even simple transformers may not accurately capture the dynamics of our natural system

## 4 Discussion

The risk of flooding of low-lying coastal land is closely related to high water levels in adjacent coastal waters. Peaks in water surface elevation occur as rare events when

meteorological conditions combine with spring high tides. While tidal fluctuations in water level can be reproduced and forecast with high precision and confidence, non-tidal effects (residuals) are not routinely forecasted. We implemented three ML models in addition to a regression model used in prior work, showing their superior performance in forecasting these anomalous extreme events in the Petaluma River estuary close to its confluence with San Francisco Bay. The deviations of ensemble model output from observed residuals are compared in Figure 8, showing that LSTM ensemble performs the best in handling the extreme event and associated high variability in January 2023.

After evaluating our results it becomes clear that both MLPs and LSTMs can perform well for this forecasting problem. Transformers on the other hand fall short and demonstrate their inadequacy for this particular problem. This is consistent with previous work that has also demonstrated the same. Comparing with Sanah et. al.[23], we see a similar discrepancy in performance happens between MLPs/LSTMs and Transformers. The authors found that modeling Nonstationary forcing (Earth Systems with dynamics/statistical properties that change over time) with Transformers resulted in far less robust and accurate forecasts than their MLP and LSTM counterparts, similar to what we see here. Additionally, our results reflect an earlier study by Zeng et. al.[24] that investigated the general use of Transformers for time series forecasting. Here the authors found that transformers performed significantly worse than even their linear counterparts, similar to what we found. Our conclusion mirrors the authors in that self-attention is inadequate in its ability to correctly attribute temporal dependencies throughout a time series.

Looking now at the "successful" models, MLPs did quite well overall though did have significant variance, particularly on the extreme values. The success of this model is likely due to its modularity and ease of integration with the hindcast we used, as well as its ability to model the nonlinearities in our data. The variance on the extreme values is due to the MLPs weak ability to fit to out-of-distribution points. This is balanced however by the simplicity and cost-effectiveness of the model, as we see the ensemble mean generally performs well in fitting to the extreme events. It would be quite easy to scale the number of ensemble members even higher for more consistent results in regards to extreme values. This would also maintain the model's generalizability over the development of some bespoke solution for the extreme values in the Petaluma River.

Best of all, the LSTM performs well overall with low variance and high-fidelity to the actual water level in the Petaluma River. This was expected as LSTMs were specifically developed for time series data and forecasting. The success of the model is primarily due to its direct inclusion of previous temporal information from the time series which reinforces the model's predictive capability for weak or out-of-distribution points. This results in far more confident predictions for extreme values. However, we note that the LSTMs tended to slightly under-predict the extreme values, which is a concern in the context of social preparedness. Under-predicting water level elevation and thus flood risk may result in the absence of government or community response during a flood. On the other hand, over-predicting extreme water levels and flood risk incurs unnecessary economic and political costs (e.g., closing a major roadway when it was not necessary). In implementing forecasts, government and community will need to account for error in the method when making operational decisions (as they do with existing inferior forecasts).

A key component of model skill, as we have demonstrated here, is ensembling. While the overall skill of an ensemble is close to its individual members, their value for this system lies in forecasting outliers. Model ensembles tend to forecast outliers with more consideration for the surrounding time series and are regularized to the extent where strong shifts in any one feature do not devolve into radically large changes in water level residual. This regularization is expected behavior as shown by Abati et. al.[16] and Wang et. al.[17].

Additionally, Sanah et. al.[23] show that ensembling also improves upon the physical modeling of Earth/Ocean Systems. This aligns well with what we see here as our ensembles tend to localize large SHAP values within extreme values in the forecast. This is perhaps easiest to see in LSTM results, fig. 7, where we can see that the SHAP values in the second half of the figures align with the elevated portion of the water level forecast. The LSTM ensemble has a clear drop off in SHAP values at the end of the elevated water level residual while the single LSTM has a much slower drop off in importance. As in Sanah et. al.[23] we find that ensembles improve our confidence in the ability of each

model to represent physical concepts. The low standard deviations in our model ensembles, fig. 8, indicate a strong ability to learn underlying physical concepts rather than relying on coincidental single model results.

While we are satisfied with the performance of our models, we also set out to explicitly target and model the physics of the Petaluma River through the correlations between hydrodynamic and atmospheric factors. We do not impose specific constraints and regularization to explicitly target these, but through inclusion of physically correlated features we can model the underlying physics of the system. We know this is true looking at our SHAP values, fig. 7, where we see concurrence between multiple features that combine to explain extreme water level events. Specifically, we see higher water levels during high river flow, onshore winds in the bay, wind-driven surges along the open coast, and low atmospheric pressure. Capturing physical mechanisms is important as it allows our models to be robust to future changes in the system.

With climate change and sea-level rise, the water level in estuaries like the Petaluma River becomes increasingly non-stationary (less likely to follow statistical norms). Similar work by Lai et. al.[25] and Yik et. al.[26] show that physics-based modeling of conditions related to climate change and non-stationary systems does not perform well, emphasizing our choice of physically-correlated, local features. This allows us to establish a local model that is robust enough to capture short-term, local conditions. Considering that our framework allows for the continuous update of the model parameters, this should also ensure that our models account for any changes in climate over time - in other words, our model can learn from new data and adapt as the effect of the physical processes on local water levels evolve. Sanah et. al.[23] use a similar framework for the modeling of physical properties in the Atlantic Meridional Overturning Circulation (AMOC). Their results also mirror ours in that both MLPs and LSTMs are capable of capturing the underlying physical dependencies of these Earth systems, particularly through the use of ensembling. The ensembling encourages overall adherence to the underlying physics by mitigating individual ensemble members who deviate on harder-to-forecast points.

To ensure that our models do what we want, we need to understand how the models are making predictions, which we explore through SHAP values. SHAP and AI explainability as a whole are extremely important to oceanography, highlighted by Sonnewald et. al.[27] and Sanah et. al.[23], in order to increase trust in ML models as well as improve their generalizability. Sonnewald et. al.[27] notes that explainability methods like SHAP enable oceanographers to understand how the physical dynamics of ocean and river systems interact and identify shortcomings in the models themselves, as we do here. By identifying the strengths and shortcomings of models using XAI techniques we can then adjust models so that they are both more generalizable and trustworthy.

With this in mind, looking at our SHAP values more closely (fig. 7) we conclude that the success of the MLPs and LSTMs over the Transformers is due to better mapping of the features (physical drivers) and their importance to the residual water levels in Petaluma River. This correctness is represented by more localized importance around the extreme water levels and a smoother transition between states (negative to positive importance and vice versa) aligning with our knowledge of the physical system where extreme values are the product of numerous overlapping factors. The high-frequency variability in SHAP values for Transformer models indicates a lack of fidelity to lower frequency forcing as illustrated by SHAP values for the other models tested. This smooth transition in SHAP values for LSTM and MLP represents the buildup and passing of storm systems and their effect on the water level. Again, this behavior is mirrored in the work by Sanah et. al.[23] where not only do Transformers struggle to capture the same dependencies as MLPs and LSTMs, but the resultant SHAP values also demonstrate a disconnect with underlying known physics in the models. This reinforces our confidence in both MLPs and LSTMs as a models suitable for time-varying phenomena like extreme events in complex earth systems.

Overall, we are more confident in MLPs and LSTMs due to the smoothness in their SHAP values which aligns with our expectation of the system where there is more of a buildup to extreme values. Transformers on the other hand demonstrate erratic behavior that is only emphasized by the extreme shifts in their SHAP values. Furthermore, ensembling does positively impact the physical modeling capabilities of a model, as shown by Sanah et. al.[23].

The ultimate question is whether these models are useful. Our motivation is to support

adequate preparation for potential floods adjacent to the Petaluma River, including transport routes like State Highway-37. Our immediate aim is to forecast water levels in the River. Adequate preparation depends on an accurate and timely forecast of an extreme water level event that will lead to flooding of the nearby area. While MLPs will almost always lead to adequate preparation, the tendency of the MLPs to over-predict extreme values means that while we will be prepared, we will likely spend significant time and energy preparing for floods that never happen. This is a significant problem as Highway 37 is a major commuter highway and shutting it down unnecessarily will cause community backlash and distrust. The LSTMs are the most accurate models in their ability to forecast extreme values. However, they have the tendency to slightly under-predict extreme values, with the associated risk of being under-prepared for a flood that could lead to the destruction of property and the loss of life. Managers will need to weigh these risks of negative or positive flood warning errors within the context of their specific decisions and authorities.

Regardless, both of these model types are improvements over doing nothing (i.e., using simple tidal predictions) and over regression approaches. Both the LSTM and MLP ensembles consistently forecast when the residual is positive and with an acceptable amount of error. Nevertheless, the models have room for significant improvement, which will occur as more data becomes available including additional extreme events. The inclusion of more data, whether it be an increased training window or the addition of more physically correlated features, will lead to improvements in all the model types we tested. Petaluma River (actually a tributary estuary) is a representative case study for coastal flooding, and it can be expected that our models will also perform well in other estuary/bay sites where sufficient data are available. This approach could enable accurate forecasting for many sites prone to coastal flooding with modest data needs (a single pressure sensor at the vulnerable site).

## 5 Conclusion

In this study we examine the effectiveness of MLPs, LSTMs, and Transformers at predicting extreme water level values in the Petaluma River. We additionally include ensembles of each model type and examine their overall trustworthiness through the use of Shapley Values. We show that both MLPs and LSTMs perform very well and are suitable improvements over prior Regression models. In contrast, we find that Transformers have limited utility for this task, but encourage a broader investigation into their adaptation to this field.

Addressing extreme values and their explainability directly, we found that both MLP and LSTM ensemble models could accurately forecast extremes. Further, the explanations of factors accounting for extreme events matched physical understanding and were validated by domain experts. This highlights the importance of including explainability when targeting anomalous or extreme values in predictions as it enhances model trust and enables both the scientific and local communities to take steps to address these extremes.

Lastly, we also demonstrate the usefulness in regularizing neural network based architectures like our MLPs, LSTMs, and Transformers. Through this regularization we achieved not just more stable predictions, but also more inclusive predictions in regards to the overall water level in the Petaluma River. Ensembling enables us to quantify the uncertainty in our models, building on the trust we can associate with them and also enabling us to verify the physics involved in the system.

Overall, we find that the inclusion of nonlinearities is useful in water level forecasting. These nonlinearities can also be intelligently added and explained using simple methods like SHAP. We find that ensembling is an effective way to regularize the latent space of these models, effectively including outliers and extreme values in the ensembles predictions which can then also be verified using SHAP.

**Acknowledgments**
T his work was supported by California Department of Transport (CalTrans), through the National Center for Sustainable Transportation at UC Davis, and by California State Parks. Additional support was from the University of California Davis. We are grateful for the assistance of our colleagues in the Coastal Oceanography Group at Bodega Marine Laboratory who conducted the fieldwork and managed the data to allow this study.

**Data availability**
A ll the data used in this study is publicly available. Wind and water flow data can be found through NOAA (https://www.ndbc.noaa.gov/) and USGS (https://dashboard.waterdata.usgs.gov/app/nwd/en/) respectively. The data and sensor id for the Petaluma River can be requested by contacting the Bodega Marine Laboratory.

**References**

[1] Thirumalaiah K and Deo M C 1998 *Journal of Hydrologic Engineering* **3** 26–32 URL https://ascelibrary.org/doi/abs/10.1061/%28ASCE%291084-0699%281998%293%3A1%2826%29

[2] Campolo M, Andreussi P and Soldati A 1999 *Water Resources Research* **35** 1191–1197 (*Preprint* https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/1998WR900086) URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1998WR900086

[3] 2019 accessed 25 Feb 2026 URL https://dot.ca.gov/-/media/dot-media/district-4/documents/d4-environmental-docs/4q320-sr37-flood-reduction/4q320-mrn-sr37-ded-master-2023august23-508-a11y.pdf

[4] US Geological Survey 2026 The national map accessed 25 Feb 2026 URL https://apps.nationalmap.gov/viewer/

[5] Hochreiter S and Schmidhuber J 1997 *Neural Computation* **9** 1735–1780 ISSN 0899-7667 (*Preprint* https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf) URL https://doi.org/10.1162/neco.1997.9.8.1735

[6] Orozco López E, Kaplan D and Linhoss A 2024 *Water Resources Research* **60** e2023WR036337 e2023WR036337 2023WR036337 (*Preprint* https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2023WR036337) URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023WR036337

[7] Xu S and Huang W 2010 *Artificial Neural Network Model Application on Long Term Water Level Predictions of South Florida Coastal Waters*

[8] Vizi Z, Batki B, Rátki L, Szalánczi S, Fehervary I, Kozák P and Kiss T 2023 *Environmental Sciences Europe* **35**

[9] Lundberg S M and Contributors S 2025 *SHAP Documentation* URL https://shap.readthedocs.io/en/latest/

[10] Lee S, Purushwalkam S, Cogswell M, Crandall D and Batra D 2015 *Why M Heads are Better than One: Training a Diverse Ensemble of Deep Networks* (*Preprint* 1511.06314) URL https://arxiv.org/abs/1511.06314

[11] Leutbecher M and Palmer T 2008 *Journal of Computational Physics* **227** 3515–3539 ISSN 0021-9991 predicting weather, climate and extreme events URL https://www.sciencedirect.com/science/article/pii/S0021999107000812

[12] Renate Hagedorn F J D R and Palmer T 2005 *Tellus A: Dynamic Meteorology and Oceanography* **57** 219–233 (*Preprint* https://doi.org/10.3402/tellusa.v57i3.14657) URL https://doi.org/10.3402/tellusa.v57i3.14657

[13] European Centre for Medium-Range Weather Forecasts (ECMWF) 2025 *Development of the European Multi-Model Ensemble System for Seasonal to Interannual Forecasts* URL https://www.ecmwf.int/en/forecasts/dataset/development-european-multimodel-ensemble-system-seasonal-interannual

[14] Moradi R, Berangi R and Minaei B 2020 *Artificial Intelligence Review* **53** 3947–3986 ISSN 1573-7462 URL https://doi.org/10.1007/s10462-019-09784-7

[15] Breiman L 1996 *Machine Learning* **24** 123–140 URL https://link.springer.com/article/10.1007/BF00058655

[16] Abati D, Porrello A, Calderara S and Cucchiara R 2019 *Latent Space Autoregression for Novelty Detection* (*Preprint* 1807.01653) URL https://arxiv.org/abs/1807.01653

[17] Wang N, Chen C, Xie Y and Ma L 2020 *Brain Tumor Anomaly Detection via Latent Regularized Adversarial Network* (*Preprint* 2007.04734) URL https://arxiv.org/abs/2007.04734

[18] Bowman W 2025 *UTide: Unified Tidal Analysis and Prediction* URL https://github.com/wesleybowman/UTide

[19] Developers S L 2025 *PolynomialFeatures - Scikit-Learn Documentation* URL https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

[20] Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H and Zhang W 2021 *Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting* (*Preprint* 2012.07436) URL https://arxiv.org/abs/2012.07436

[21] Hegazy K, Mahoney M W and Erichson N B 2025 *Powerformer: A Transformer with Weighted Causal Attention for Time-series Forecasting* (*Preprint* 2502.06151) URL https://arxiv.org/abs/2502.06151

[22] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2023 *Attention Is All You Need* (*Preprint* 1706.03762) URL https://arxiv.org/abs/1706.03762

[23] Suri S and Sonnewald M 2026 *ESS Open Archive* **2026** (*Preprint* https://essopenarchive.org/doi/pdf/10.22541/essoar.176894678.89831689/v1) URL https://essopenarchive.org/doi/abs/10.22541/essoar.176894678.89831689/v1

[24] Zeng A, Chen M, Zhang L and Xu Q 2022 *Are Transformers Effective for Time Series Forecasting?* (*Preprint* 2205.13504) URL https://arxiv.org/abs/2205.13504

[25] Lai C Y, Hassanzadeh P, Sheshadri A, Sonnewald M, Ferrari R and Balaji V 2025 *Annual Reviews Condensed Matter Physics* **16** 343–365

[26] Yik W, Sonnewald M, Clare M C A and Lguensat R 2023 Southern ocean dynamics under climate change: New knowledge through physics-guided machine learning (*Preprint* 2310.13916) URL https://arxiv.org/abs/2310.13916

[27] Sonnewald M, Lguensat R, Jones D C, Dueben P D, Brajard J and Balaji V 2021 *Environmental Research Letters* **16** 073008 URL https://doi.org/10.1088/1748-9326/ac0eb0