



# The Role of Machine Learning, Incorporating Wind Data, to Predict Coastal Wave Overtopping

Michael McGlade<sup>1</sup>✉, Nieves G. Valiente<sup>1</sup>, Jennifer Brown<sup>2</sup>, Christopher Stokes<sup>2</sup>, and Timothy Poate<sup>1</sup>

<sup>1</sup> University of Plymouth, Plymouth PL4 8AA, UK  
michael.mcglade@plymouth.ac.uk

<sup>2</sup> National Oceanography Centre, Southampton SO14 3ZH, UK

**Abstract.** Anthropogenic climate change is increasing wave overtopping frequencies. Reliably predicting overtopping is important for minimising casualties and economic losses. Empirical formulas, such as EurOtop, can provide accurate overtopping predictions based on physical model experiments. EurOtop requires nested regional process-based models to provide the input of nearshore coastal conditions. These models are computationally expensive, restricting their application in operational forecasting. EurOtop excludes wind speed and direction; significant variables influencing overtopping. This study explores the role of training random forests, using annual overtopping observations, to predict overtopping at two study sites: Dawlish and Penzance (UK). We compare the performance of random forests with a EurOtop based model, while investigating the performance of including hourly measured wind speed and direction. The random forests estimated overtopping and non-overtopping with approximately 80% and 95% accuracy, respectively. Including wind speed and direction enhanced the random forests' predictive performance ( $R^2 = > 0.70$ ,  $F1 = > 0.80$ ). Random forest outperformed the predictive precision and accuracy made by EurOtop at Dawlish ( $F1 = 0.81$ ,  $F1 = 0.65$ , respectively) and Penzance ( $F1 = 0.86$ ,  $F1 = 0.07$ , respectively). Random forests generated predictions rapidly ( $< 4$  s) and generalised well, demonstrating great potential for implementation in different overtopping hotspots worldwide, provided overtopping observation availability.

**Keywords:** Machine-Learning · Overtopping · EurOtop · Flooding

## 1 Introduction

By 2050, approximately 300 million individuals globally may experience high-frequency coastal flooding [1]. Globally aggregated annual overtopping hours have increased by almost 50% over the last two decades [2]. Predicting wave overtopping along seawall defences is challenging given the number of variables influencing flooding. Such variables include wind speed ( $U_{10}$ ), wind direction ( $U_{10}$  Dir), significant wave height ( $H_s$ ), freeboard, wave direction ( $D_m$ ), and mean period ( $T_m$ ). These variables interact and

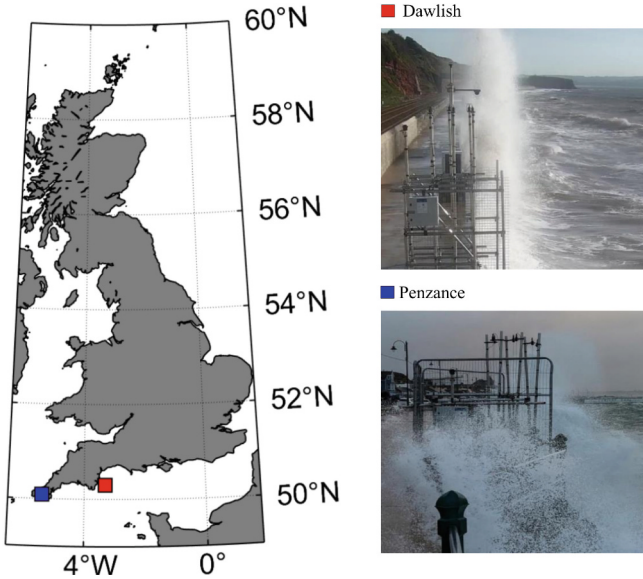
can change significantly spatially and temporally. The EurOtop manual estimates overtopping by considering various conditions, such as hydraulic variables (e.g.  $H_s$ ) and structural parameters (e.g. crest freeboard), and then calculates overtopping discharge by accounting for sloped and vertical seawall structures [3]. Many process-based models rely on using the EurOtop manual for estimating overtopping. These models often incur errors due to inadequate spatiotemporal resolutions, bathymetry, forcing data, and seawall geometries [4]. Wind effects are also oversimplified due to insufficient data, and although several studies have demonstrated the effects of wind on wave overtopping [5, 6], current widely used empirical parameterisations still neglect these. Machine learning, specifically random forests, may offer new insights to predict overtopping rapidly and reliably. Random forests consist of multiple decision-tree ensembles, analysing different sections of the training dataset, which then formulate together a final prediction [7]. Training random forests over several storms may reliably predict overtopping, derived from learning correlations from historical overtopping events [8]. Including  $U_{10}$  and  $U_{10}$  Dir in the random forests may also improve such predictive performances. This study examines the accuracy and precision of training random forests, using field observational overtopping data, along with the corresponding wave and wind conditions, to predict overtopping occurrence and frequency (number of overtopping waves per 10 min) in Dawlish and Penzance. We compare the importance of accounting for wind conditions and compare the random forest performance with EurOtop predictions following the process-based approach used in a forecasting system called SWWEP OWWL [9] which is operational across the study region.

## 2 Methods

### 2.1 Study Sites and Overtopping Measurements

Wave overtopping was investigated in Dawlish ( $50^{\circ}35'01''$  N,  $3^{\circ}27'52''$  W) and Penzance ( $50^{\circ}07'10''$  N,  $5^{\circ}32'15''$  W), southwest England (Fig. 1). Both locations experience frequent overtopping; for example, in February 2014, the Dawlish seawall collapsed, causing disruption amounting to £50 million [10]. The Dawlish seawall reaches 5.64 m above Ordnance Datum, and the Penzance wall reaches 3.7 m. The seawall structures differ, with Dawlish consisting of a recurved section for deflecting waves, while Penzance is slightly curved ( $35^{\circ}$ ). The WireWall apparatus, developed by the National Oceanography Centre (NOC), was permanently deployed at Dawlish and Penzance seawalls from March 10<sup>th</sup> 2021 to March 17<sup>th</sup> 2022, and November 16<sup>th</sup> 2021 to March 15<sup>th</sup> 2022, respectively [11, 12]. The systems were programmed to run for 6 h every tide centered over predicted high tide. At Dawlish, two WireWall systems were adjacently deployed to create a cross-shore transect (50 cm apart); rig 1 at the seawall crest and rig 2 landward close to the railway line. At Penzance, two WireWall systems were deployed with along-shore separation, rig 1 was exposed to SE & SW winds (Eastern seawall location), and rig 2 was more sheltered, exposed to SW winds (western seawall location) (Fig. 1). The corresponding wind conditions (i.e.,  $U_{10}$  and  $U_{10}$  Dir) were recorded using local weather stations, presenting wind data every 10 min. Wave buoys recorded the wave conditions, computing corresponding wave conditions every 30 min. WaveRadar REX, a marine

monitoring system, recorded the water level every 10-min and 15-min for Dawlish and Penzance, respectively.



**Fig. 1.** Study sites for Dawlish (red) and Penzance (blue) illustrating Wirewall setup.

## 2.2 Machine Learning Development

Wind, wave, water level, and WireWall overtopping observations were processed in 10-min intervals. WireWall overtopping occurrence was classified binary (yes/no) as well as using the mean number of overtopping events computed in 10-min windows. A binary and regression random forest classifier estimated overtopping occurrence and frequency. We selected a random forest given its efficiency in handling limited data, its ability to rank feature variable importance, and its capability of rapidly generating predictions through its ensemble decision-tree structure. The data was partitioned into testing (20%) and training (80%) ratios. 10-fold cross-validation determined this optimised test-to-train ratio. Systematic grid searches analysed the hyperparameter tuning metrics. For the binary model, the threshold was adjusted to optimise the precision and recall. The optimal threshold configuration was 0.43. L2 regularisation addressed overfitting; selected given the few features. The random forest consisted of 100 decision trees, max depth = 10, and sample splits equaling = 10. Variable importance was determined by SHapley Additive exPlanations (SHAP) [13].

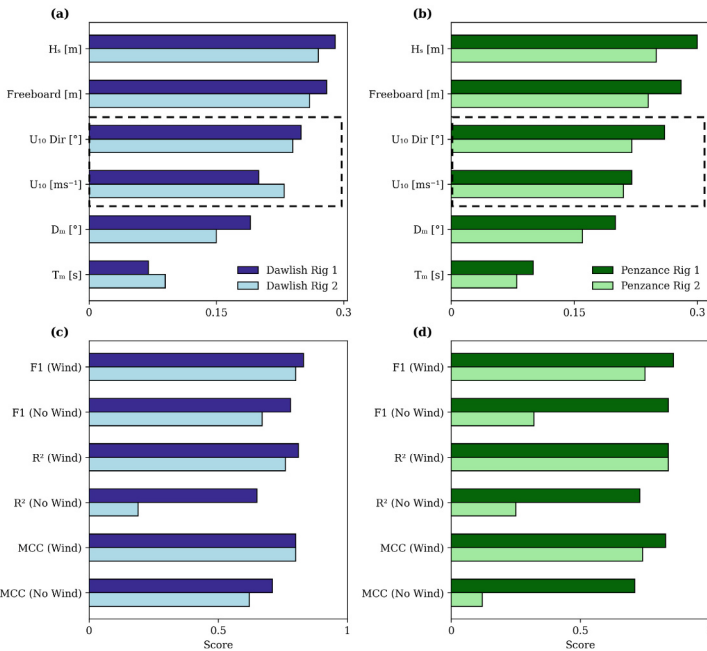
EurOtop predictions were generated using the same wave and tidal observations as the machine learning model. These conditions were transformed into the coast using the parametric breaker dissipation model, transforming wave conditions from the buoy to Dawlish and Penzance seawalls, with adjustments for foreshore slope, toe mound effects, and impulsive breaking, following the approach of SWEEP OWWL [14]. EurOtop overtopping volumes exceeding 0.1L/s/m were classified as overtopping (1), and

discharge volumes below this threshold as non-overtopping (0), reflecting pedestrian safety thresholds in the EurOtop manual.

### 3 Results and Discussion

#### 3.1 Variables Influencing Overtopping

Random forests demonstrated strong predictive performances for estimating overtopping occurrences and frequency in Dawlish and Penzance ( $F1 > 0.80$ ,  $R^2 > 0.70$ ) (Fig. 2c,d).  $H_s$  and freeboard are the most important variables influencing overtopping (Fig. 2a,b). Larger  $H_s$  corresponds to more energetic wave conditions, increasing wave runup, and the probability of waves reaching the seawall crest [15]. Lower freeboard conditions reduce the barrier height that waves must overcome, increasing overtopping likelihood.  $U_{10}$  and  $U_{10}$  Dir are also important (Fig. 2a, b). Excluding  $U_{10}$  and  $U_{10}$  Dir decreased the random forest performance for estimating overtopping occurrences, particularly at rig 2 in Penzance ( $F1 = 39\%$  without wind; Fig. 2d). Omitting wind variables reduces the random forest’s ability to estimate the number of overtopping events in Dawlish ( $R^2 = 0.54$  and  $R^2 = 0.19$ , respectively) (Fig. 2c).



**Fig. 2.** Random forest variable importance metrics for (a) Dawlish and (b) Penzance comparing with and without wind variables for (c) Dawlish and (d) Penzance.

#### 3.2 Comparing Machine Learning with EurOtop

The random forest yielded high-performance metrics for estimating overtopping in Dawlish and Penzance ( $F1 > 0.80$ ), demonstrating low instances of false positives and

negatives (precision > 80%) (Table 1). Random forests indicated high performances for estimating non-overtopping events ( $F1 = 0.98$ ). Calculating a single prediction by the random forests was under 4 s. The SWEEP OWWL demonstrated low to moderate performance for estimating overtopping in Dawlish ( $F1 = 0.65$ ) with very low predictive performances in Penzance ( $F1 = 0.07$ ). The high accuracy from the random forest could be attributed to its inherent design, aiming to reduce predictive model bias and variance. Random forests sample the training dataset with replacement, meaning each decision tree specialises in different dataset sections. Random forests rank feature variables of importance (Fig. 1), focusing less on redundant unnecessary features, reducing model complexity, bias, and miss-predictions. However, the random forests did indicate several false positives and negatives which require further analysis.

The SWEEP OWWL bases its predictions for overtopping discharge using EurOtop, which is a neural network configuration. Neural networks are powerful machine learners, robust to high noise and variance. The likelihood of SWEEP OWWL's poor performance likely attributed more to the challenges of addressing data imbalances, rather than its configurations of not handling noisy data. Across the WireWall deployment, overtopping was recorded at approximately 2–3%, whereas non-overtopping was >97%, indicating significant class imbalances. Neural networks require extensive training, significantly more than random forests, given the number of neurons, hidden layers, and bias values requiring adjustment for neural networks. This neural network may not be trained enough on the minority overtopping class, likely causing this poor performance. The high performance on the non-overtopping class supports this, indicating that because many cases were non-overtopping, this sufficed the training requirements for the neural network; however, the neural network is likely overfitting the data to non-overtopping correlations, causing the low predictive performance for the overtopping class. Sweep OWWL does not consider wind parameters in its overtopping predictions, with many studies emphasising the importance of influencing wind, which influences  $H_s$ , alters  $D_m$ , and affects the landward transport of sea spray.

**Table 1.** SWEEP OWWL performance against the random forest for Dawlish and Penzance.

SWEEP OWWL					Random Forest		
Location	Phase	F1	Recall	Precision	F1	Recall	Precision
Dawlish	Overtopping	0.65	0.73	0.59	0.81	0.77	0.86
Penzance	Overtopping	0.07	0.04	1	0.86	0.82	0.89

## 4 Conclusion

This study demonstrates the high predictive performance of random forest models to estimate overtopping. Incorporating wind variables significantly improved random forest predictions. This study calls for further research to contribute additional guidance to EurOtop to improve the prediction of wind-related conditions. Our study reveals that

random forests can significantly outperform SWEEP OWWL, which bases its predictions for overtopping discharge on EurOtop. Random forest demonstrated great potential for implementation in different hotspots worldwide, given its strong generalisation performance for both study locations, providing overtopping observations are available. Ensuring appropriate generalisation to different locations, this random forest would need training to predict overtopping on various seawall structures.

## References

1. Kulp A, Strauss H (2019) New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding. *Nat Commun* 1(1):4835–4844
2. Almar R et al (2021) A global analysis of extreme coastal water levels with implications for potential coastal overtopping. *Nat Commun* 12(3775):3770–3775
3. EurOtop. <http://www.overtopping-manual.com>. Accessed 11 Dec 2024
4. Buccino MD, Leo A, Tuozzo L, Calabrese D (2023) Wave overtopping of a vertical seawall in a surf zone: a joint analysis of numerical and laboratory data. *Ocean Eng* 1(288):116114–116120
5. Van Der Van Gent MRA, Bijl RJ, Wolters G, Wüthrich D (2024) The maximum influence of wind on wave overtopping at seawalls with crest elements. *Emerald Publishing* 1(9999):1007–1021
6. Di Leo A, Dentale F, Buccino M, Tuozzo S, Pugliese C (2022) Numerical analysis of wind effect on wave overtopping on a vertical seawall. *Water* 14(23):3875–3891
7. Rong G et al (2020) Rainfall induced landslide susceptibility mapping based on Bayesian optimized random forest and gradient boosting decision-tree models - a case study of Shuicheng County. *China Water* 12(11):3055–3066
8. Naeem S, Ali A, Anam S, Ahmed M (2023) An unsupervised machine learning algorithm: comprehensive review. *Int J Comput Digit Syst* 1(1):1–12
9. Stokes K, Poate T, Masselink G, King E, Saulter A, Ely N (2021) Forecasting coastal overtopping at engineered and naturally defended coastlines. *Coast Eng* 1(164):103820–103827
10. Dawson D, Shaw J, Gehrels WR (2016) Sea-level rise impacts on transport infrastructure: the notorious case of the coastal railway line at Dawlish, England. *J Transp Geograph* 1(51):97–109
11. Brown J et al (2023) Coastal wave overtopping: new nowcast and monitoring technologies. *Coast Eng Proc* 37(1):1–15
12. Yelland J et al (2024) A system for in-situ, wave-by-wave measurements of the speed and volume of coastal overtopping. *Commun Eng* 2(9):1–9
13. Deb D, Smith R (2021) Applications of random forest and SHAP tree explainer in exploring spatial (in)justice to aid urban planning. *Int J Geo-Inf* 10(9):629
14. Janssen T, Battjes J (2007) A note on wave energy dissipation over steep beaches. *Coast Eng* 54(9):711–7116
15. Salauddin M, Pearson M (2019) Wave overtopping and toe scouring at a plain vertical seawall with shingle foreshore: a physical model study. *Ocean Eng* 1(171):286–299

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

