Dual-Head GRU for Predicting Dangerous Sea Conditions and Wave Heights from Buoy Data

Paula Pešić

August 11, 2025

Abstract

This mini-project explores forecasting dangerous sea conditions using data from the Irish weather buoy network. A dual-head GRU neural network with attention pooling was trained on 72-hour sequences of 47 engineered features to predict whether a dangerous sea state will occur in the next six hours and forecast the mean wave height over the same period. Data from buoys M2 and M3 were used for training and in-domain testing, while buoy M5 was held out for generalization testing. The model achieved high AUC (0.936–0.974) and strong regression accuracy (R^2 up to 0.933), showing good performance both in-domain and under spatial domain shift.

1 Introduction

Motivation Forecasting dangerous sea conditions is important for ship safety, coastal management, and public alerts \cite{caires2005global, bidlot2002intercomparison}. Early warnings can help vessels change course and allow authorities to act in time. This project relies on historical buoy data (wave height, wind speed, and weather measurements) to predict hazardous wave conditions six hours ahead. The data contains many variables, is often noisy, and has missing values. Because of this, it needs thorough preprocessing to fill gaps, standardize features, and capture time patterns before building the prediction model.

Approach A *dual-head* gated recurrent unit (GRU) based neural network with attention pooling is used to addresses two related tasks: (1) **Binary classification** — predicting if a dangerous sea condition will occur within the next $H = 6$ hours, using seasonal 90th percentile significant wave height (WVHT) thresholds; (2) **Regression** — forecasting the average WVHT (m) over the same horizon.

The model processes **72 consecutive hourly timesteps** ($T = 72$) with **47 standardized features** per timestep. It includes meteorological and oceanographic variables, cyclical time encodings, and missingness indicators. Attention pooling allows the network to focus on the most relevant parts of the sequence, while Focal Loss addresses class imbalance in rare extreme events. This architecture supports safety-critical binary detection and continuous forecasting for planning.

Research Questions

- Can a GRU, trained on historical buoy sensor data, accurately forecast dangerous wave events?
- How well does the trained model generalize to buoy data from locations not seen during training, under domain shift?

2 Dataset and Preprocessing

Source and Scope The dataset used is the *Weather Buoy Network* [1], containing hourly meteorological and oceanographic measurements from the Irish moored weather buoy network [3]. Stations include M1, M2, M3, M4, M5, M6, FS1, M4-Archive, and Belmullet-AMETS, with variables such as:

- Atmospheric pressure (mbar), air temperature (°C), dew point (°C), sea temperature (°C), relative humidity (%)
- Wind speed and gust (knots), wind direction ()
- Significant wave height (WVHT, m), wave period (s), maximum wave height (Hmax, m), and mean wave direction ()

Missing or unavailable data are recorded as `NaN` or `-999`. The raw CSV (`Buoy_raw.csv`, 63.49 MB) contains **613,392** hourly records from **9 stations**, spanning **2001-02-06** to **2017-11-28**.

Data Quality and Missingness Initial exploration showed moderate missingness in several variables (e.g., 9–12% for wind speed, WVHT, and wave period) and very high missingness (>78%) for directional wave metrics (`MeanWaveDirection`, `Hmax`). To retain information from incomplete records, binary missingness masks and gap-length (_delta) features were generated before imputation.

Station Coverage and Danger Rates Coverage (observed vs. expected hours) varied by buoy, from 64% (FS1) to 92% (M1). Danger rates[2] ranged from 9.4% (M4) to 15.9% (M2). Stations were selected based on coverage, missingness, and diversity in sea patterns:

- **Main buoy (M2)** — primary training, validation, and in-domain test data.
- **Extra training buoy (M3)** — added to broaden training conditions.
- **Generalization buoy (M5)** — out-of-domain evaluation. M5 was the nearest buoy to M2 that met the minimum distance (200 km) and maximum correlation ($\rho \leq 0.85$) criteria, with actual WVHT correlation $\rho \approx 0.76$.

Preprocessing Pipeline The main preprocessing steps were:

1. Cleaning and type conversion; removing latitude/longitude to prevent location leakage.
2. Replacing placeholders (`-999` → NaN) and angular encoding of wind and wave directions into sine–cosine components.
3. Continuous hourly reindexing; interpolation for gaps $\leq$12 hours, monthly medians for longer gaps (per buoy and per month).
4. Computing seasonal (per-month) 90th percentile WVHT thresholds from raw wave heights for danger labeling.
5. Feature engineering: relative wave height (normalized to threshold), cyclical time encodings, and missingness mask/_delta features.

[1] `https://www.kaggle.com/datasets/mahdiehhajian/weather-buoy-network`

[2] Proportion of hours where WVHT exceeded the seasonal 90th percentile threshold for that buoy.

Target Construction From each buoy, overlapping 72-hour windows were extracted with stride = 3 hours. Targets were:

- $y_{\text{cls}} = 1$ if any continuous 2-hour period in the next 6 hours exceeded the seasonal WVHT threshold.
- y_{reg} = mean WVHT over the same 6-hour horizon (meters).

Splitting and Scaling For in-domain evaluation, M2+M3 sequences were split chronologically into 70% train, 15% validation, and 15% test, ensuring no overlapping timestamps. M5 was processed identically for generalization testing. A `StandardScaler` was fit on all training-domain sequences (M2 and M3) only, then applied to validation, test, and generalization sets to avoid leakage.

Final Processed Dataset Statistics After cleaning, engineering, and sequence generation:

Table 1: Final sequence counts and danger rates.

Split	Sequences	Danger rate (%)	Shape $(N, 72, 47)$
Train (M2+M3)	64,416	9.71	(64,416, 72, 47)
Validation	13,804	6.71	(13,804, 72, 47)
Test	13,804	9.44	(13,804, 72, 47)
Generalization (M5)	38,287	8.05	(38,287, 72, 47)

All processed datasets were saved in compressed `.npz` format with aligned feature names, timestamps, and thresholds for reproducibility.

3 Methodology

3.1 Task Setup and Inputs

The model operates on **72-hour sequences** ($T = 72$ time steps), each with **47 standardized features** aligned across all splits (Train/Val/Test/Gen). Two targets are defined over the following $H = 6$ hours:

1. **Binary classification:** Predict whether dangerous wave conditions will occur, defined using seasonal 90th percentile WVHT thresholds.
2. **Regression:** Forecast the average significant wave height (WVHT, m) over the same horizon.

Recorded danger rates differ by split: Train 9.71% (926 danger / 9543 total), Val 6.71% (926 danger / 13 804 total), Test 9.44% (1 303 danger / 13 804 total), and Gen (M5) 8.05% (3 082 danger / 38 287 total). Before proceeding, the data was checked to ensure correct feature alignment ($input\_dim = 47$) and to confirm there is *no* future leakage in inputs or timestamp overlap between splits or buoys.

3.2 Architecture: GRU with Attention Pooling

The network consists of:

- **GRU backbone:** Two stacked GRU layers (`hidden_size`=64, `num_layers`=2) with dropout $p = 0.3$ applied between GRU layers.
- **Attention pooling:** A learnable attention mechanism assigns weights to each timestep, producing a context vector that emphasizes the most relevant parts of the input sequence. An additional dropout layer ($p = 0.3$) is applied to the pooled context.
- **Dual output heads:**
  - **Classification head:** Linear layer outputting logits for binary danger prediction.
  - **Regression head:** Linear layer predicting the mean WVHT (m).

3.3 Loss Functions and Class Imbalance Handling

The total training loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Focal}} + 0.5 \cdot \mathcal{L}_{\text{MSE}}$$

where:

- $\mathcal{L}_{\text{Focal}}$: Focal Loss ($\alpha = 1$, $\gamma = 2$) mitigates class imbalance by focusing on hard, rare danger cases. A `pos_weight` of ≈ 9.30 was computed for reference, but not directly applied in the Focal Loss.
- $\mathcal{L}_{\text{MSE}}$: Mean Squared Error for regression.

Mini-batches are balanced using a `WeightedRandomSampler`.

3.4 Optimization and Early Stopping

- **Optimizer:** Adam with learning rate 10^{-3}.
- **Scheduler:** `ReduceLROnPlateau` (mode=max, factor=0.5, patience=3), stepped on a combined score:

  $$\text{Score} = 0.4 \cdot \text{Val AUC} + 0.6 \cdot \text{EMA(Gen AUC)}$$

  (EMA smoothing factor = 0.3).
- **Regularization:** Gradient clipping (`max_norm` = 1.0).
- **Early stopping:** Patience 5 on the combined score.

3.5 Evaluation Tools

Evaluation integrates standard metrics and additional analysis for robustness:

- **Metrics:** ROC AUC, F1, precision, recall for classification; RMSE, MAE, and R^2 for regression.
- **Calibration:** Temperature scaling (optimal temperature ≈ 0.7944) applied to logits on the validation set.
- **Uncertainty:** Monte Carlo Dropout ($N = 20$ passes) at inference to estimate epistemic uncertainty and produce mean predictions.

- **Feature importance:** Permutation-based drop in AUC for each feature.
- **Visualization:** ROC and PR curves, confusion matrices, probability distributions, predicted vs. true regression plots, and reliability diagrams.

4 Experimental Results

This section reports the performance of the Dual-Head GRU with attention pooling on the in-domain (main buoy) test set and the out-of-domain (generalization buoy),using temperature-scaled probabilities and MC Dropout mean predictions ($N = 20$ passes). Classification metrics are computed at the F1-optimal threshold from the validation split ($\tau = 0.504$). Regression RMSE values are reported in meters.

Table 2: Performance summary (calibrated, MC Dropout, $\tau = 0.504$). RMSE is in meters.

Split	AUC	Precision	Recall	F1	RMSE	R^2
Val	0.974	0.722	0.725	0.723	0.431	0.933
Test	0.936	0.499	0.708	0.586	0.876	0.518
Gen	0.969	0.578	0.838	0.684	0.441	0.851

4.1 Classification

The classification head achieved consistently high AUC across all splits: 0.973 (Val), 0.936 (Test), and 0.969 (Gen). This indicates strong ability to rank dangerous and safe conditions. AUPRC values were 0.783, 0.649, and 0.656 respectively. This shows the effect of class imbalance. Hazard recall was high on all splits (0.757 Val, 0.748 Test, 0.864 Gen), which is critical for reliable warning systems.

Figure 2 shows confusion matrices, predicted probability distributions, and PR curves for each split. Recall remains high even under domain shift, but precision is lower for Test and Gen due to more false positives. Probability separation between safe and danger classes is clearest in Val, with greater overlap in Test and Gen, consistent with the PR curves.

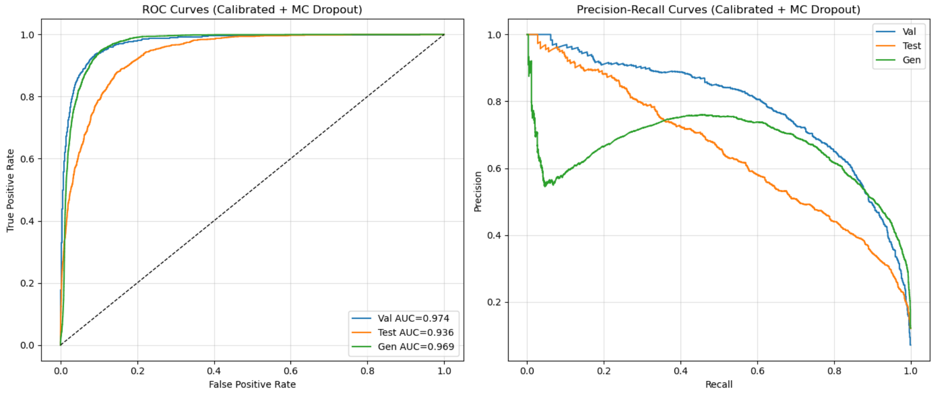

Figure 1: ROC and PR curves (calibrated + MC Dropout) for Val, Test, and Gen splits.

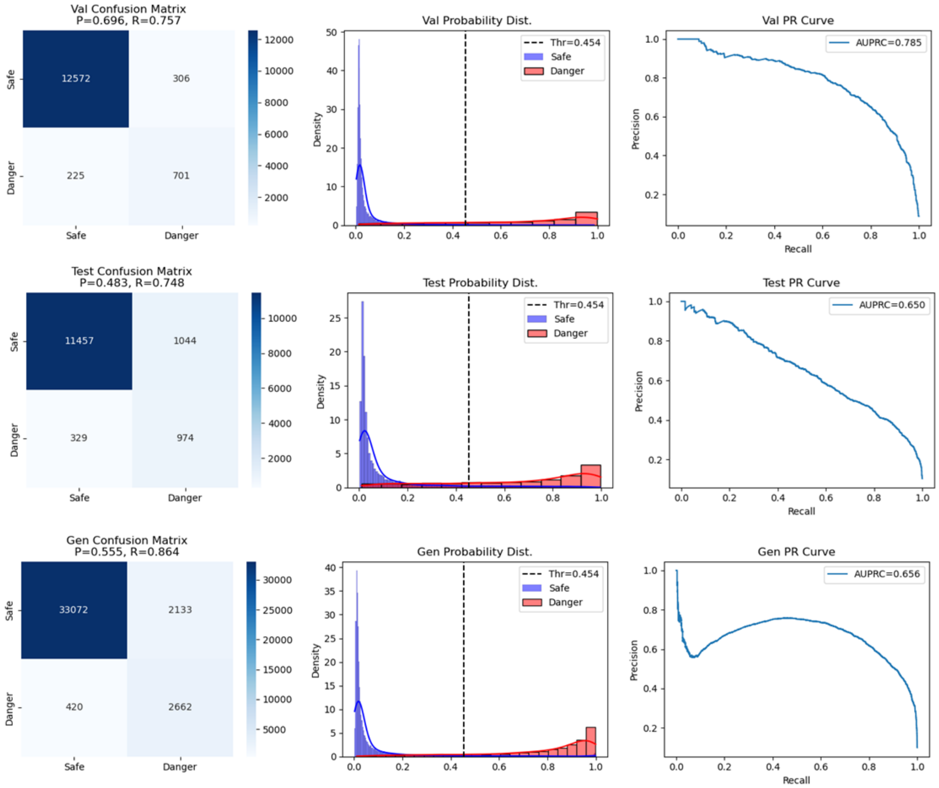

Figure 2: Classification results for Val (top row), Test (middle), and Gen (bottom). Each row shows the confusion matrix (left), predicted probability distribution with decision threshold $\tau = 0.454$ (middle), and PR curve with AUPRC (right).

4.2 Regression

For the 6-hour mean WVHT forecast, the model performed well on the validation ($R^2 = 0.933$) and generalization ($R^2 = 0.851$) splits, with RMSE values of 0.43 m and 0.44 m, respectively. The test set showed lower accuracy ($R^2 = 0.518$, RMSE 0.88 m), likely due to differences in wave height distributions compared to the training data.

Mean bias was small for Val (+0.123 m) and Gen (+0.097 m), but larger for Test (+0.460 m). This indicates a tendency to overpredict wave heights in the out-of-sample in-domain case. Error plots (Fig. 3) show mild bias. In the Test split, errors increase for larger wave heights. The error distributions (Fig. 4) are narrow and centered close to zero for Val and Gen, but wider and more skewed for Test, matching the lower R^2.

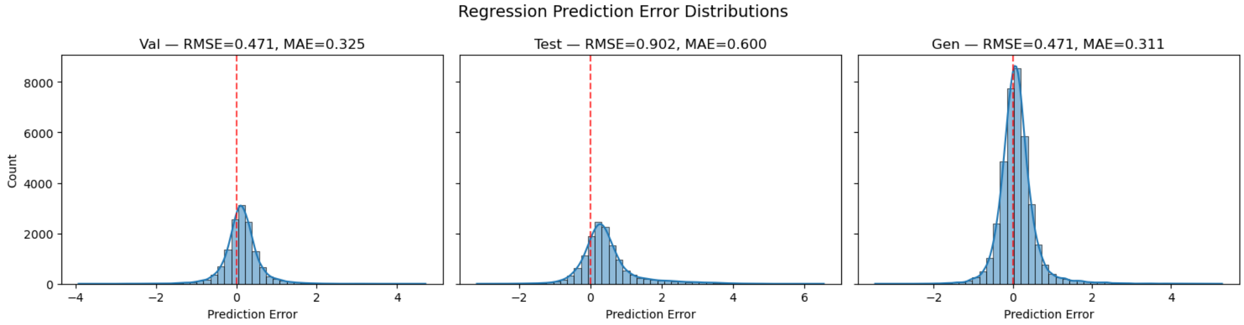

Figure 3: Predicted–true WVHT scatterplots with error residuals per split.

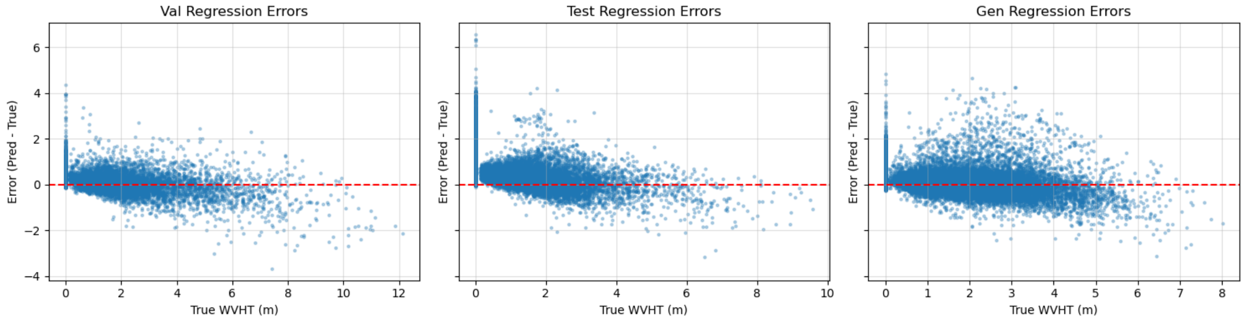

Figure 4: Regression error distributions with RMSE/MAE per split.

4.3 Calibration and Uncertainty

Probability calibration is evaluated using temperature scaling and the Brier score. The optimal scaling temperature on the validation set was $T \approx 0.7944$. After scaling, the Brier scores were 0.0288 (Val), 0.0655 (Test), and 0.0469 (Gen). This suggests good calibration in-domain and a small drop under domain shift.

Figure 5 shows reliability diagrams for all three splits. Val and Gen are well-calibrated, while Test shows slight underconfidence.

Uncertainty was also estimated using Monte Carlo (MC) Dropout. Across splits, higher uncertainty was linked to borderline wave height cases near the danger threshold, while confident predictions were typically very safe or very dangerous. This pattern held in both in-domain and out-of-domain data.

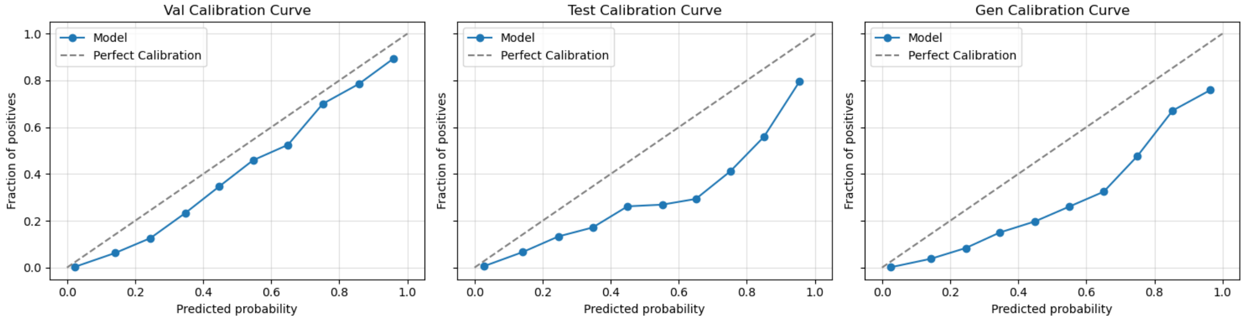

Figure 5: Calibration curves after temperature scaling: fraction of positives vs. predicted probability for Val (left), Test (middle), and Gen (right).

4.4 Feature Importance

To assess which input variables most strongly drive the hazard classification head of the GRU with attention pooling, permutation feature importance was computed separately on the validation (in-domain) and generalization (out-of-domain) sets.

The validation set results (Table 3) is a skewed importance distribution. Two features (F14 and F3) dominate, with AUC drops of 0.1233 and 0.1089 respectively, corresponding to 100% and 88.3% of the maximum observed effect. The top five features together account for over 40% of the total cumulative importance (right panel, Fig. 6). This means that the model's hazard forecasts are concentrated on a small subset of sensor variables, and not on low-signal features.

In the generalization set (Table 4), F14 remained the dominant predictor, while other top features shifted slightly. This suggests that the model retains core hazard cues but adapts to the conditions of a new buoy.

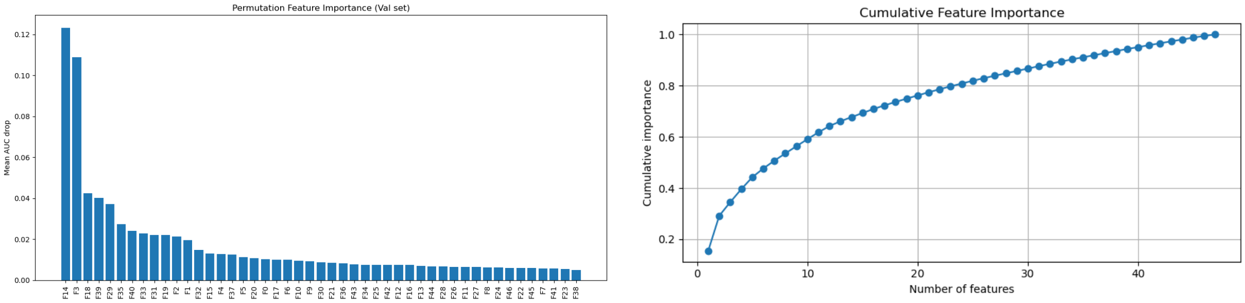

Figure 6: Permutation importance (validation set, left) and cumulative importance (right).

Table 3: Top 5 features by relative permutation importance (validation set).

Feature	Mean AUC drop	Relative to max (%)
F14	0.1233	100.0
F3	0.1089	88.3
F18	0.0425	34.5
F39	0.0402	32.6
F29	0.0371	30.1

Table 4: Top 5 features by relative permutation importance (generalization set, 20% sample).

Feature	Mean AUC drop	Relative to max (%)
F14	0.2112	100.0
F18	0.0226	10.7
F33	0.0198	9.4
F3	0.0195	9.2
F32	0.0194	9.2

5 Discussion

The dual-head GRU with attention pooling performed well for classification in both in-domain and out-of-domain cases (AUC: 0.936–0.974), with hazard recall always above 0.70. This means the model can reliably detect dangerous sea states. The main drawback was lower precision when tested on a different buoy, which led to more false alarms. This could be improved by using adaptive thresholds or buoy-specific calibration.

For regression, accuracy was high on the validation and generalization sets (\$R2̂ ¿ 0.85\$) but much lower on the in-domain test set (\$R2̂ = 0.52\$). This drop is likely due to differences in wave height patterns and wind–wave relationships. Feature importance analysis showed that only a few physical variables provided most of the predictive power, suggesting the model could still work well with fewer sensors.

6 Conclusion

A GRU-based dual-task model with attention pooling can forecast dangerous wave events and short-term wave heights from buoy data, while generalizing well to an unseen buoy location. Although performance drops under some domain shifts, the approach is still practical for maritime hazard forecasting. Future improvements could include adding external environmental data, applying buoy-specific calibration, and exploring architectures with longer temporal context to improve robustness.

References

[1] Caires, S., and Sterl, A., "Global wave height trends and variability," *Bulletin of the American Meteorological Society*, vol. 86, no. 4, pp. 495–500, 2005.

[2] Bidlot, J.R., et al., "Intercomparison of the performance of operational ocean wave forecasting systems with buoy data," *Weather and Forecasting*, vol. 17, no. 2, pp. 287–310, 2002.

[3] Marine Institute, "Weather Buoy Network," *data.gov.ie*, 2017. Available: `https://data.gov.ie/publisher/marine-institute`.