Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine-learning prediction of late-stage saprolegniosis risk in freshwater aquaculture from water-quality data

Petra Camilla Lindholm-Lehto

petra.lindholm-lehto@luke.fi

Natural Resources Institute Finland

Harri Vehviläinen

Natural Resources Institute Finland

Jaana Jurvansuu

Alvoa

Article

Keywords: Aquaculture; Saprolegnia, landlocked salmon, water quality, machine learning, Random forest

Posted Date: November 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11387412/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Saprolegniosis, also known as water mould disease, causes substantial mortality in freshwater animals, threatening endangered species and imposing major economic losses on aquaculture. However, evidence regarding its environmental drivers remains inconsistent. We used daily aggregated water-quality data and fish saprolegniosis mortality records from a landlocked salmon hatchery to identify environmental predictors of late-stage saprolegniosis and assess their potential for early warning. A weighted random forest predicted saprolegniosis risk from water-quality data for the preceding 28 days, achieving strong predictive performance (Brier score = 0.096; log loss = 0.341; ROC–AUC = 0.989). When applied to new monitoring data, the model indicated increasing disease risk 15 days before the first observed cases of the subsequent infection period. Model interpretation associated higher late-stage saprolegniosis risk with oxygen-rich water, moderate temperatures (12–17°C), and stable organic-matter conditions, whereas high turbidity and greater variability in UV-absorbing organic matter were associated with lower risk. This transparent, data-driven framework could support the development of early-warning approaches to tackle saprolegniosis in aquaculture.

Introduction

Saprolegnia spp. are oomycetes that cause saprolegniosis in freshwater fish, amphibians, and crayfish worldwide [1, 2, 3, 4]. In aquaculture, these pathogens, particularly *S. parasitica*, affect salmonids and other cultured species, causing substantial mortality in fish and their eggs [5, 6, 7] and major economic losses to farms and hatcheries [8, 9 10]. Saprolegniosis also threatens endangered species [11, 12, 13].

Saprolegnia parasitica grows on fish through branching hyphae and disperses through water mainly via zoospores [14]. Infected fish commonly develop white or grey cotton-like growths on the skin, fins, and gills [15, 16], which impairs osmoregulation and respiratory function [2]. Advanced infections cause extensive skin damage, disrupting salt and water balance and leading to death [2, 17]. Stressed, immunocompromised, or injured fish are particularly susceptible [3, 18]. Overcrowding, concurrent infections, poor water quality, and abrupt temperature changes can increase susceptibility [19, 20]. Both high-temperature stress and rapid cooling have been implicated as factors for saprolegniosis outbreaks: Roberts [14] describes outbreaks associated with high temperatures, whereas Bly et al. [21] linked temperature declines exceeding 10°C within 24 hours to immunosuppression in pond-reared catfish. Water conditions also affect the pathogen directly. Zoospore production decreases above 20°C and below 4°C [22, 23], while humic substances and other organic matter can inhibit mycelial growth [24]. Humic substances, a complex mixture of organic compounds of decomposing organic matter, can bind calcium ions, which contribute to adhesion, encystment, and germination [25, 26]. However, it is unclear how these affect fish susceptibility and pathogen development to influence outbreaks [3].

Treatment options for saprolegniosis are limited [20]. Malachite green was effective, but it is not permitted in food-producing fish in the European Union (EC, 470/2009 and EU, 37/2010). Advanced infections remain difficult to treat; thus, good fish farming practices and suitable water quality are

therefore important for prevention [7]. Understanding how different water quality parameters and their interactions are related to disease outbreaks could help predict disease risk and allow earlier intervention [3].

Machine learning is increasingly used to predict disease outbreaks and mortality in aquaculture. Aravanis et al. [27] combined neural networks, symbolic regression, and decision trees to predict disease-associated mortality in European seabass (*Dicentrarchus labrax*) and gilt-head sea bream (*Sparus aurata*) using water temperature and husbandry data. Using seven years of surveillance data, Xie et al. [28] applied random forests to predict cryptocaryoniasis in marine fish, identifying stocking density and several water-quality variables as important predictors. Wang et al. [29] extended this approach to multiple parasitic diseases, linking disease occurrence to salinity, temperature, dissolved oxygen, and stocking density using neural networks and XGBoost.

Similar approaches have been applied to salmonids. Yilmaz et al. [30] combined water-quality measurements and bacterial profiles to predict disease occurrence in rainbow trout using multinomial logistic regression, reporting 95.7% accuracy. Öhlschuster et al. [31] developed an ensemble model to predict salmonid alphavirus transmission among Atlantic salmon farms from well-boat movements, proximity to affected farms, environmental conditions, and seasonality. The model achieved 94.4% accuracy and predicted transmission up to eight weeks before pancreas disease was officially recorded, highlighting the potential for early warning. However, the potential of machine learning to predict saprolegniosis risk from water-quality data remains underexplored.

Saprolegniosis outbreaks have become more frequent and severe over the last decade at the Natural Resources Institute Finland's (Luke) Enonkoski aquaculture facility for reasons that remain unclear. The disease threatens the endangered landlocked salmon (*Salmo salar* m. *sebago*), whose conservation depends heavily on captive broodstock and egg production. European whitefish and brown trout at the facility are also seriously affected. In this study, we developed an interpretable random forest model to predict late-stage saprolegniosis risk from water-quality data collected at Enonkoski. We combined daily aggregated water-quality measurements with infection observations, estimating infection densities on unobserved days. We aimed to identify water-quality conditions associated with disease risk and assess the model's potential for early warning.

Results

Model development and predictive performance

Because infection records were unavailable on some days due to interruptions in sampling, missing daily infection labels were estimated to reconstruct a continuous time series of infection dynamics. This yielded daily estimates of both infection presence and infection density, allowing infection risk and infection magnitude to be assessed simultaneously. The resulting series provided daily point estimates of infection density ("hard labels"), derived from observed or estimated values (Fig. 1A), and

corresponding probabilities of infection occurrence (“soft labels”), reflecting uncertainty in detection (Fig. 1B). Over the 21-month observation period, infection density exhibited three major flare-ups between May and July, with smaller peaks in October and December. Infection-free periods were most common between July and September. Although infection densities remained low from January to February 2025, estimated infection probabilities indicated persistently elevated infection risk during this period, contrasting with the same months in 2024, when multiple zero-infection days were observed and predicted probabilities were substantially lower.

Water-quality variables exhibited significant circadian variation in surface and bottom pH, pH-mV, and turbidity as determined by Kruskal–Wallis tests with Benjamini–Hochberg correction (all $p < 0.05$ after correction). Pairwise Pearson correlation analyses across time-of-day subsets showed that most variables were strongly correlated and could be represented by a single daily mean. In contrast, bottom turbidity showed weak correlations between time-of-day subsets ($|r| < 0.30$) and was therefore retained as separate daily means for morning, afternoon, evening, and night (Fig. 2).

Feature-ranking analyses consistently identified surface temperature as a dominant predictor across modelling approaches (Fig. S1), although agreement in the exact ranking of predictors was limited across scenarios (Fig. S2). Applying a 14-day temporal-spacing constraint reduced redundancy among closely related predictors while preserving dominant temporal signals, such as longer-term temperature summaries (Fig. S3). The final predictor set spanned all major water-quality variables, with strongest representation from dissolved surface oxygen and bottom temperature, and included a mixture of lagged values, rolling variability, and rolling means, reflecting multiple temporal scales of environmental influence on infection risk.

Random Forest (RF), LightGBM (LGB), and XGBoost (XGB) models were evaluated against a constant baseline and a temporal persistence baseline (t − 1). All models substantially outperformed the constant baseline in both Brier score and log-loss, demonstrating that environmental predictors captured meaningful temporal structure in infection dynamics. Overall calibration did not consistently exceed the persistence baseline, reflecting strong temporal autocorrelation; however, for hard labels, RF and XGB achieved lower log-loss than persistence, indicating sharper probabilistic predictions. Model performance was largely insensitive to the number of predictors (Top-K = 10–50), indicating stable feature rankings (Fig. S4). Variance-based weighting slightly reduced performance for hard labels but improved calibration for soft labels. Across learners, RF showed the most stable and interpretable performance and was therefore selected as the final model. After optimisation, the RF achieved a mean Brier score of 0.096 ± 0.037 and a log-loss of 0.341 ± 0.194, improving substantially over the constant baseline (Brier = 0.229; log-loss = 0.668). Discrimination was high (ROC-AUC = 0.989 ± 0.002; PR-AUC = 0.995 ± 0.004). Predictive performance remained stable across probability thresholds, with maximal F1 values 0.95 around a threshold of 0.33 (Fig. 3A). At the conventional midpoint threshold (P(infection) = 0.5), the model correctly classified 342 of 375 days, missing 16 infection events (Fig. 3B).

When applied to an independent test period (September–October 2025), the model produced realistic short-term forecasts of infection risk (Fig. 3C). Predicted infection probabilities remained low (< 0.3) during the initial infection-free period (1–12 September), consistent with the absence of observed infections. Probabilities then increased steadily and exceeded the calibrated decision threshold (0.5) on 15 September, coinciding with the first detected infected fish. Notably, infection probability began to rise approximately 15 days before the first observation, indicating that environmentally driven changes preceded visible infection. This lead time demonstrates the model's capacity to function as an early-warning indicator for routine aquaculture disease monitoring.

Water-quality predictors of late-stage saprolegniosis risk

Based on SHAP values the 20 most influential features contributing to infection probability included surface and bottom measurements of dissolved oxygen, temperature, turbidity, UV254 (total and filtered), and surface measurements of pH and nitrate nitrogen (NO_3-N) (Fig. 4A). Block permutation importance analysis produced a close matching ranking. The largest increases in Brier score occurred when surface and bottom dissolved oxygen, temperature, and UV254 features were permuted, confirming their predictive significance under serial dependence (Fig. 4B). Both SHAP and permutation analyses identified dissolved oxygen (surface 11-day lag; bottom 9-day mean) and bottom temperature (13-day mean; 3-day lag) as the strongest predictors.

The SHAP beeswarm revealed that low dissolved oxygen values decreased infection probability, whereas higher oxygen concentrations were associated with increased risk. Conversely, for bottom temperature, very high values reduced infection probability while moderately lower values supported it. Other consistently important features included surface dissolved oxygen 28-day mean, surface turbidity (6-day lag and 28-day standard deviation), UV254 total (surface 18-day standard deviation and 2-day mean; bottom 6-day lag), surface temperature (13- and 28-day means), and bottom temperature 28-day standard deviation.

The SHAP dependence plots further showed that low turbidity values (6-day lag) tended to support infection, whereas high variation (28-day standard deviation) reduced it. Similarly, for UV254 total, high short-term variation (18-day standard deviation) or low recent values (2–6 days prior) reduced infection probability, while more stable conditions appeared favourable for infection. The overall agreement between SHAP and permutation results indicates that the model's dominant predictors are both temporally stable and biologically interpretable rather than artefacts of autocorrelation.

Effects of predictors and their interactions on infection probability

To understand how environmental conditions measured by water quality parameters shaped infection probability, quantile-based SHAP dependence plots were examined for the 16 most influential features

(Fig. 5). Each plot showed how variation in a single predictor affected the model's infection probability while accounting for interactions with related variables.

Surface dissolved oxygen at an 11-day lag had a clear transition point near 8.8 mg L^{-1}. The infection probability sharply decreased below and above it (Fig. 5A). The similar shape of the 27-day lag feature indicates that longer-term oxygen conditions followed a comparable pattern, though typically at slightly higher values (Fig. 5A). Bottom dissolved oxygen (9-day mean) also showed that increase between 7.8 and 9.3 mg L^{-1} supported infection (Fig. 5B). The peak at 8–9 mg L^{-1} coincided with periods of high UV254t variation (18-day standard deviation). This co-occurrence implies that infection risk was elevated during short-term oxygen enrichment events coupled with increased organic-matter variability in the water column at the time of the lake full cycle. On the surface, 28-day mean of dissolved oxygen interacted strongly with surface temperature (Fig. 5E). When oxygen exceeded 9 mg L^{-1} and temperature ranged between 10–15°C, infection probability was highest. However, low oxygen (< 9 mg L^{-1}) suppressed infection even under otherwise favorable temperatures, highlighting oxygen's limiting role.

Bottom temperature features revealed a clear thermal optimum for infection. The 13-day rolling mean indicated maximum infection support between 11°C and 17°C, with infection probability declining sharply above 18°C (Fig. 5C). The 3-day lagged bottom temperature confirmed this pattern: infection probability increased around 15°C but fell rapidly at higher temperatures (Fig. 5D). The 28-day mean temperature further showed that sustained warming above 10°C supported infection, emphasizing that *S. parasitica* proliferation required a longer period of stable, moderate warmth rather than transient spikes (Fig. 5D). Finally, surface temperature (28-day mean) showed an infection optimum at 12–18°C, beyond which infection probability declined (Fig. 5H). The combined effects of oxygen and temperature revealed that infection was suppressed when water exceeded 18°C and oxygen dropped below 9 mg L^{-1}, whereas moderate temperature (10–15°C) and high oxygen (over 10 mg L^{-1}) together promoted infection. Low temperatures still supported infections as the oxygen levels were also high.

Turbidity and UV254t features acted in the opposite direction. Moderate turbidity (1.8–2.0 NTU; 6-day lag) supported infection, whereas high turbidity values (over 2.4 NTU) reduced infection probability, particularly when bottom temperatures (3-day lag) exceeded 15°C (Fig. 5F). Similarly, for UV254t (surface 18-day standard deviation), low to moderate values (0.05–0.30 abs/m) with previously (17-day lag) high bottom temperature (over 20°C) supported infection. On the other hand, very high variation in values had no apparent effect irrespective of temperature (Fig. 5G).

Overall, the SHAP dependence plots indicated that *Saprolegnia* infections were most likely under moderate (10–17°C), oxygen-rich (> 9 mg L^{-1}), low turbidity, and low variability in UV254t, with risk dramatically decreasing under high temperature (above 20 ° C), low oxygen, high organic matter conditions (turbidity > 2.4 NTU or extreme variability in UV254t).

Discussion

Our findings show that water-quality conditions during the preceding 28 days provided predictive information about late-stage saprolegniosis risk. The model represents late-stage saprolegniosis because labels were based on dying or dead fish with visible water-mould growth. Random Forest was selected after comparison with LightGBM and XGBoost. It showed high discrimination (ROC–AUC = 0.98; PR–AUC = 0.995), although infection-positive days accounted for 81% of the modelling data. The Random Forest model's Brier score did not improve on the previous-day persistence baseline, indicating that recent infection history remained a strong benchmark. Nevertheless, when applied to new data, the model indicated increasing risk 15 days before the first observed cases of a subsequent infection period.

Late-stage saprolegniosis risk was associated with temperature, oxygen, turbidity, and organic-matter conditions over the preceding days. These conditions may affect both the pathogen and the fish's defences against infection. Fish skin mucus forms a protective barrier, and its composition can change with developmental stage, water pH, infection, and stress [32, 33]. Handling can increase stress and damage the skin, increasing susceptibility to saprolegniosis [5, 34]. Damaged areas can provide entry points for *S. parasitica*, while mucus and its associated microbial community help maintain the skin barrier and defend against pathogens [35, 36]. These host defences may help explain why exposure to *Saprolegnia* does not always lead to disease. Pavić et al. [3] reported higher *S. parasitica* loads on injured trout than on apparently healthy fish, while Bly et al. [21] observed saprolegniosis in immunocompromised catfish exposed to zoospore levels above 5 spores mL^{-1}. Together, these findings suggest that pathogen exposure and host condition should be considered when interpreting water-quality associations.

Higher predicted risk of late-stage saprolegniosis was associated with elevated dissolved oxygen and moderate water temperatures (12–17°C). Few cases were recorded above 17°C, suggesting that the observed relationship was not simply one of increasing risk with increasing temperature. Temperature should therefore be interpreted alongside oxygen and the other water-quality predictors, rather than as an isolated driver. Temperature can affect different stages of *S. parasitica* development in different ways. Previous studies reported reduced zoospore production above 20°C [22, 23]. Mycelial growth has been reported to peak between 15 and 25°C, whereas zoospore viability and sporulation efficiency peak at around 10°C [37]. Temperature may also influence disease through effects on host susceptibility. Elevated temperatures have been associated with thermal stress and increased susceptibility to saprolegniosis [38, 39], while rapid cooling has been linked to immunosuppression in catfish [21]. Wang et al. [36] also reported that sudden cooling reduced epidermal mucous-cell numbers and caused thinning and possible detachment of the mucus layer in grass carp. Changes in host defences could therefore contribute to temperature-associated disease patterns.

Our results also suggest that the infection risk was enhanced when oxygen concentrations changed from low to moderate or high levels. Higher oxygen concentrations at cooler temperatures were associated with ongoing infections, whereas lower oxygen concentrations at warmer temperatures coincided with declining infection. At temperatures otherwise associated with higher risk, oxygen concentrations below 9 mg L^{-1} were associated with lower risk. These patterns suggest that oxygen and

temperature should be considered together, but do not establish that low oxygen directly limited infection. This differs from Bly et al. [21], who reported no direct effects of variation in oxygen or pH on saprolegniosis development under their study conditions. Only one pH-related feature ranked among the top 20 SHAP predictors, and pH was not prominent in permutation importance assessed using the Brier score. Thus, pH showed limited predictive importance within the mostly neutral to slightly acidic conditions observed in this study.

Higher turbidity and greater UV254 variability were associated with lower predicted late-stage saprolegniosis risk. These conditions occurred during July–August in both 2024 and 2025, coinciding with periods without recorded infections. Features lagged by approximately 10 days suggest that preceding water conditions provided predictive information, although this delay cannot be interpreted directly as the time required for infection or mould growth. Humic substances can inhibit mycelial growth of *Saprolegnia* [3, 24], providing one possible explanation for associations involving organic matter.

Our study extends the use of environmental data for disease-risk prediction in aquaculture [27, 28, 29]. The increase in predicted risk before observed cases also parallels the advance warning reported by Öhlschuster et al. [31] for salmonid alphavirus transmission among Atlantic salmon farms. Previous studies of salmonid disease risk incorporated bacterial profiles [30] or combined PCR results with farm and environmental data [31]. In contrast, our model generated risk estimates from automated water-quality measurements without requiring pathogen testing as a predictive input.

Limitations include partly model-estimated infection labels and the absence of management variables, such as antifungal treatments and fish handling. In addition, all data came from one facility, and advance warning was evaluated during only one subsequent infection period, limiting conclusions about reliability elsewhere. Future work could include additional water-quality measurements, such as calcium, which has been implicated in *Saprolegnia* development and infection processes [3] [26] [25]. Integrating water-quality monitoring with biological measures, such as image-based detection of fish showing visible signs of saprolegniosis [40], automated spore counts, measurements of skin and mucus properties, or fish stress-related changes, could help shift the focus towards earlier-stage infection detection.

Materials and Methods

Aquaculture description

Natural Resources Institute Finland (Luke)´s Enonkoski aquaculture infrastructure is located in southeastern part of Finland in southern Savo (62.0889, 28.9081). The aquaculture infrastructure has been established in 1989 and operates in a flow-through principle. Enonkoski aquaculture research infrastructure has an important role in the conservation of endangered stocks of valuable fish species: landlocked salmon (*Salmo salar* m. *sebago*, Vuoksi river basin), brown trout (*Salmo trutta*, Vuoksi river basin), and Saimaa arctic char (*Salvelinus alpinus*). Water is led to the farm from Lake Ylä-Enonvesi (area

11.86 km2) 3.5 L s^{-1} (13.6 m3 h^{-1}, hydraulic retention time, HRT 39 min). The inlet water is taken as a mixture of the surface (1 m) and bottom (8 m) layers. The maximum flow rate in the flow-through system is 700 L s^{-1} (permit allows 670–870 L s^{-1}). From Lake Ylä-Enonvesi, the water flows gravitationally via a 980 m tunnel to the fish farm. Coarse particles were filtered from the water. The fish were fed with Alltech Fennoaqua´s Hercules 1.2-2.5mm, and Ova 3.5 mm.

Saprolegniosis detection and infection density

From January 2024 to August 2025, experienced staff recorded fish showing visible signs of saprolegniosis during routine husbandry. Fish were identified by white or grey cotton-like growths and lesions on the skin, fins, snout, or gills [41, 42]. Records included dead fish and visibly infected, dying fish that were removed and euthanized. For each observation, we recorded the total number of fish, tank water volume, fish density, the number of affected fish, and their proportion of the total tank fish population. Infection density was calculated as the number of visibly affected fish removed divided by the tank water volume. Additional records included fish species, age class, stock origin, mean individual weight, tank identifier, tank area, and tank height. Data were extracted from farm logbooks and comprised 1,238 observation records. Landlocked salmon accounted for 69% of records, lake-dwelling brown trout (*Salmo trutta*) for 27%, and other fish for 4%. Each log entry was treated as a sampling event (Effort = 1), while days without records were treated as unsampled (Effort = 0).

Water quality measurements

Water quality was monitored using an s::can system comprising spectro::lyser V3, pH::lyser, and oxi::lyser sensors (s::can GmbH, Vienna, Austria). The sensors measured dissolved oxygen, temperature, pH, turbidity, total organic carbon (TOC), dissolved organic carbon (DOC), nitrate nitrogen (NO_3–N), and UV absorbance at 254 nm (UV254). The spectro::lyser used a 5-mm optical path and a wavelength range of 190–750 nm. Its measuring head was cleaned with compressed air before each measurement. Spectral measurements were converted into water-quality estimates using the manufacturer's software (moni::tool 4.6.7 V10.13). Both total and particle-compensated UV254 absorbance were recorded as indicators of UV-absorbing organic matter. Measurements were adjusted to local water conditions using offset calibration according to the manufacturer's instructions. Calibration used samples collected from Lake Ylä-Enonvesi across different seasons and analysed by Eurofins Lahti for suspended solids (SFS-EN 872:2005, modified), TOC (SFS-EN 1484:1997), and DOC (SFS-EN 1484:1997 RZ). The quantification limits for TOC and DOC were 1 mg L^{-1}. The pH sensor was calibrated every six months using pH 4 and pH 9 buffer solutions. Continuous monitoring sensors recorded surface and bottom measurements of temperature, pH, pH-mV, dissolved oxygen, UV254 (filtered and total), nitrate (NO_3-N), and turbidity at 10-minute intervals from March 2024 to August 2025. The raw water-quality dataset comprised 80,792 time points and 18 variables (excluding timestamps).

Estimating daily infection density using a two-part hurdle GAM

Daily infection densities were estimated from fish removal records using a two-part hurdle generalised additive model (GAM). A binomial GAM with a logit link estimated infection occurrence probability (π_i) on sampled days (Effort = 1), while a Gamma GAM with a log link estimated positive infection density (r_i), calculated as fish removed divided by sampled water volume. Both components included additive spline terms for days since monitoring began (24 basis functions) and day of week (7 cyclic basis functions), with smoothing selected by grid search. Expected daily infection density was calculated as $\mu_i = \pi_i r_i$. Parameter uncertainty was estimated by drawing from the fitted coefficient distributions to obtain prediction variances and 95% confidence intervals. Observed densities were retained where available, and model estimates filled unsampled days. Predictions below the detection threshold of one infected fish per sampling volume for at least two consecutive days were set to zero. Binary infection-presence indicators provided hard labels, while probabilities from the logistic component provided soft labels. Daily infection densities, estimated variances, and soft labels were used in subsequent environmental modelling.

Feature processing and selection

Surface and bottom measurement were treated as distinct features throughout processing and selection. Water-quality measurements were recorded at 10-minute intervals and averaged by calendar day. Before averaging, time-of-day differences were assessed across morning, afternoon, evening, and night using Kruskal–Wallis tests with Benjamini–Hochberg correction. Correlations between daily values for these periods were then examined for each variable. Missing daily or time-of-day means were imputed before cross-validation using a multivariate iterative imputer with Bayesian Ridge regression, a maximum of 20 iterations, and random seed 0.

From 21 environmental variables, we generated 1,764 temporal predictors comprising lagged values and rolling means and standard deviations over 1–28 days, using only preceding observations. Random Forest, LightGBM, and XGBoost were used to identify informative features. Analyses considered both binary infection-presence labels, defined as infection density > 0, and continuous occurrence-probability labels, with and without inverse-variance weighting. Native tree-model feature importances were normalised to [0, 1] within each learner and combined across folds and modelling scenarios. Features were ranked using mean importance and majority votes. To reduce temporal redundancy, a ± 14-day spacing rule was applied within each variable and feature type (lag, rolling mean, or rolling standard deviation). After selecting a feature, candidates within this interval were excluded. Precomputed feature sets containing 10–50 predictors, in increments of five, were then evaluated with and without sample weighting.

Model selection

Models were evaluated using three-fold expanding-window time-series cross-validation. Observations were ordered by date, and each model was trained on earlier observations and evaluated on the immediately following block, without shuffling or a gap between training and testing. Each test block contained approximately one-quarter of the eligible observations, while the training set expanded across folds. Three folds were used because additional splits would leave some folds without zero-infection observations. The same splits were used for all models, feature-set sizes, and weighting scenarios. Feature lists were fixed before this evaluation rather than selected separately within each training fold.

Sample weights were calculated as $u_i=[\max(v_i,10^{-6}) + 10^{-6}]^{-1}$ and normalised to a mean of one, where vi is the GAM-estimated infection-density variance. Here, (v_i) was the GAM-estimated infection-density variance for binary-label analyses and the occurrence-probability variance for continuous-label analyses. Non-finite variances were replaced using the median calculated with np.nanmedian. Weights were applied during both training and scoring.

Model selection prioritised the Brier score and log loss to assess probabilistic predictive performance. Feature-selection stability was assessed using Jaccard overlap and Kendall's τ, averaged across fold pairs. Agreement among models and scenarios was visualised using heatmaps of normalised importance values and reciprocal ranks. ROC–AUC and PR–AUC were not used for model ranking.

Final Random Forest model

The final model was a calibrated Random Forest classifier using 50 selected lagged and rolling features with the uncertainty-based sample weights to predict late-stage saprolegniosis risk. Random Forest hyperparameters were optimised during model selection and fixed for final model fitting. The selected parameters were n_estimators = 1000, max_depth = 3, min_samples_leaf = 4, min_samples_split = 2, max_features = 0.5, bootstrap = True, and criterion = "gini". The forest was fitted on the first 80% of the chronologically ordered development series, and its probabilities were calibrated using the remaining 20%. This calibration subset was excluded from fitting the forest but had contributed to earlier preprocessing, target estimation, and feature selection. Calibration was performed once after model selection, not within each cross-validation fold. The calibrated model was then applied to new data to assess whether predicted risk increased before the first observed cases of a subsequent infection period. A separate forest with the same parameters was fitted on all development data for SHAP and permutation analyses.

Model interpretation

SHapley Additive exPlanations (SHAP) were used to describe each predictor's contribution to model predictions. Features were ranked by their mean absolute SHAP values, and the stability of these rankings across folds was assessed using Kendall's τ. For dependence plots, feature values were divided into 20 quantile bins, and the mean SHAP value was plotted for each bin. A secondary feature selected using shap.utils.approximate_interactions was used to colour each plot, with the colour scale restricted to its 5th–95th percentiles. Predictor importance was also assessed using block permutation

importance. Values were permuted in contiguous 14-day blocks to preserve temporal structure within blocks while assessing changes in model performance.

Python libraries, data sets and scripts created

All analyses were conducted in Python 3.11 using pandas 2.3.1(6), NumPy 2.2.5(7), scikit-learn 1.5.0 (8), statsmodels 0.14 (9), matplotlib 3.10.0 (10), SciPy 1.14 (11), SHAP 0.48.0 (12), pyGAM 0.9.1 (13), LightGBM 4.6.0 (14), and XGBoost 3.0.5 (15). All scripts and datasets are publicly available in the GitHub repository (jjurvans/saprolegnia-infection-ml-2025).

Declarations

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contribution

P.L-L. and H.V. designed the study and arranged all practicalities in performing the measurements and monitoring the fish husbandry. P.L-L. applied for funding and acted as a project manager. J.J. provided the code and performed the data analysis. All authors (P.L-L., H.V., J.J.) wrote the manuscript and reviewed it before submission.

Acknowledgement

The experiment was funded by the European Union through the European regional Fund (EAKR) The Finland 2021–2027 innovation and skills in Finland programme, municipality of Enonkoski, and Hanka-Taimen Oy. We would like to thank the staff at the Natural Resources Institute Finland´s Enonkoski fish farm. Special thanks also to Hanka-Taimen Oy´s Pankakoski and Sulkava fish farms for data collection and collaboration.

Data Availability

All data is available at https://github.com/jjurvans/saprolegnia-infection-ml-2025.

References

1. van den Berg, A. H., McLaggan, D., Diéguez-Uribeondo, J. & van West, P. The impact of the water moulds *Saprolegnia diclina* and *Saprolegnia parasitica* on natural ecosystems and the aquaculture

industry. *Fungal Biol. Rev*. 27, 33–42. (2013). doi.org/10.1016/j.fbr. 2013.05.001

2. van West, P. *Saprolegnia parasitica*, an oomycete pathogen with a fishy appetite: new challenges for an old problem. *Mycologist* **20**, 99–104. doi.org/10.1016/j.mycol.2006.06.004 (2006).
3. Pavić, D. et al. Tracing the oomycete pathogen *Saprolegnia parasitica* in aquaculture and the environment. *Sci. Rep.* **12**, 16646. doi.org/10.1038/s41598-022-16553-0 (2022).
4. Ortega-López, P. & Diéguez-Uribeondo, J. How Saprolegniales became successful parasites. *PLoS Pathog*. **22**, e1014359. org/10.1371/journal.ppat.1014359 (2026).
5. Bruno, D., van West, P. & Beakes, G. Saprolegnia and other Oomycetes in *Fish diseases and disorders, viral, bacterial and fungal infections*, (ed. Woo, P., Bruno, D.) 669–720Wallingford, UK,. (2011).
6. Zahran, E., Hafez, E. H., Hossain, F. M. A., Elhadidy, M. & Shaheen, A. A. Saprolegniosis in Nile Tilapia: Identification, molecular characterization, and phylogenetic analysis of two novel pathogenic *Saprolegnia* Strains. *J. Aquat. Anim. Health*. **29**, 43–49. doi.org/10.1080/08997659.2016.1259691 (2017).
7. Ali, S. E., Gamil, A. A. A., Skaar, I., Charo-Karisa, H. & Ø. E. & Efficacy and safety of boric acid as a preventive treatment against Saprolegnia infection in Nile tilapia (*Oreochromis niloticus*). *Sci. Rep.* **9**, 18013. doi.org/10.1038/s41598-019-54534-y (2019).
8. Duan, J. et al. A review: factors affecting outbreaks of Saprolegniosis on aquatic animals. *Israeli J. Aquacult.* **70**, 11 (2018).
9. Mostafa, A. A. F., Al-Askar, A. A. & Yassin, M. T. Anti-saprolegnia potency of some plant extracts against *Saprolegnia diclina*, the causative agent of saprolengiasis. *Saudi J. Biol. Sci.* **27**, 1482–1487. doi.org/10.1016/j.sjbs.2020.04.008 (2020).
10. Songe, M. M. et al. *Saprolegnia diclina* IIIA and *S. parasitica* employ different infection strategies when colonizing eggs of Atlantic salmon, *Salmo salar* L. *J. Fish. Dise*. **39**, 343–352. doi.org/10.1111/jfd.12368 (2016).
11. Eronen, A., Janhunen, M., Hyvärinen, P., Kortet, R. & Karvonen, A. The effects of hybridization and parasite infection on the survival and behaviour of endangered landlocked salmon subject to predation—Implications for genetic rescue. *Evolut Appl.* **17**, e70056. doi.org/10.1111/eva.70056 (2024).
12. Korkea-aho, T. L., Viljamaa-Dirks, S. & Holopainen, R. Luke vesiviljely, Savon Taimen Oy, Voimalohi Oy. Vesihomeen ajallinen esiintyminen ja estomenetelmien optimointi kalanviljelylaitoksilla (In Finnish). *Finnish Food Authority* 4257/05.08/2021. (2021).
13. Korkea-aho, T., Wiklund, T., Engblom, C., Vainikka, A. & Viljamaa-Dirks, S. Detection and quantification of the oomycete *Saprolegnia parasitica* in aquaculture environments. *Microorganisms* **10**, 2186. doi.org/10.3390/microorganisms10112186 (2022).
14. Roberts, R. J. *Fish pathology* (4th ed.). Blackwell Publishing, ISBN: 978-1-4443-3282-7. (2012). doi.org/10.1002/97811 18222942
15. Yanong, R. P. Fungal diseases of fish. *Vet. Clin. N Am. Exotic Anim. Pract.* **6**, 377–400. doi.org/10.1016/s1094-9194(03)00005-7 (2003).

16. Elgendy, M. Y. et al. Onion (*Allium cepa*) improves Nile tilapia (*Oreochromis niloticus*) resistance to saprolegniasis (*Saprolegnia parasitica*) and reduces immunosuppressive effects of cadmium. *Aquacult. Int.* **31** (3), 1457–1481. doi.org/10.1007/s10499-022-01035-x1 (2023).
17. Engblom, C. Identification and genetic characterization of *Saprolegnia parasitica*, isolated from farmed and wild fish in Finland. *J. Fish. Dis.* **46**, 849–860. 10.1111/jfd.13795 (2023).
18. Gozlan, R. E. et al. Current ecological understanding of fungal-like pathogens of fish: what lies beneath? *Front. Microbiol.* **5**, 62. doi.org/10.3389/fmicb.2014.00062 (2014).
19. Howe, G. E. & Stehly, G. R. Experimental infection of rainbow trout with Saprolegnia parasitica. *J. Aquatic Animal Health* 10, 397–404. (1998). doi.org/10.1577/1548-8667(1998)010⟨0397: EIORTW⟩2.0.CO;2.
20. Lindholm-Lehto, P. C. & Pylkkö, P. Saprolegniosis in aquaculture and how to control it? Aqua. Fish Fisheries ;4:e2200. (2024). https://doi.org/10.1002/aff2.200. (2024).
21. Bly, J. E., Lawson, L. A., Szalai, A. J. & Clem, L. W. Environmental factors affecting outbreaks of winter saprolegniosis in channel catfish, *Ictalurus punctatus* (Rafinesque). *J. Fish. Dis.* **16**, 541–549. doi.org/10.1111/j.1365-2761.1993.tb00890.x (1993).
22. Diéguez-Uribeondo, J. Adaptation to parasitism of some animal pathogenic *Saprolegniaceae*. Comprehensive summaries of Uppsala Dissertations from the Faculty of Science and Technology 122. Acta Universitatis Upsaliensis (1995).
23. Kitancharoen, N., Yuasa, K. & Hatai, K. Effects of pH and temperature on growth of Saprolegnia diclina and S. parasitica isolated from various sources. *Mycosci* **37**, 385–390. doi.org/10.1007/BF02460994 (1996).
24. Meinelt, T. et al. Reduction in vegetative growth of the water mold *Saprolegnia parasitica* (Coker) by humic substance of different qualities. *Aquat. Toxicol.* **83**, 93–103. doi.org/10.1016/j.aquatox.2007.03.013 (2007).
25. Burr, A. W. & Beakes, G. W. Characterization of zoospore and cyst surface structure in saprophytic and fish pathogenic *Saprolegnia* species (Oomycete fungal protists). *Protoplasma* **181**, 142–163. doi.org/10.1007/BF01666393 (1994).
26. Rezinciuc, S., Sandoval-Sierra, J. V., Ruiz-León, Y., van West, P. & Diéguez-Uribeondo, J. Specialized attachment structure of the fish pathogenic oomycete *Saprolegnia parasitica*. *PLoS ONE*. **13**, e0190361. https://doi.org/10.1371/journal.pone.0190361 (2018).
27. Aravanis, T., Hatzilygeroudis, I. & Spiliopoulos, G. Ensemble modelling for predicting fish mortality. *Appl. Sci.* **14**, 6540. https://doi.org/10.3390/app14156540 (2024).
28. Xie, X. et al. Machine learning-driven early warning system for cryptocaryoniasis in marine aquaculture. *Parasit. Vectors*. **18** 10.1186/s13071-025-07124-z (2025).
29. Wang, X. et al. MarParPred: A Shiny web application for predicting multi-parasite disease risk in marine fish using machine learning. *Aquacult. Eng.* **114**, 0144–8609. doi.org/10.1016/j.aquaeng.2026.102735 (2026).

30. Yilmaz, M., Çakir, M., Oral, O., Oral, M. A. & Arslan, T. Using machine learning technique for disease outbreak prediction in rainbow trout (*Oncorhynchus mykiss*) farms. *Aquacult. Res.* **53**, 6721–6732. https://doi.org/10.1111/are.16140 (2022).
31. Öhlschuster, M. et al. On the prediction of SAV *transmission* among Norwegian aquaculture sites. *Prev. Vet. Med.* **224**, 106095 (2024).
32. Ellis, A. E. Innate host defense mechanisms of fish against viruses and bacteria. *Dev. Comp. Immunol.* **25**, 827–839 (2001).
33. Alzate-Díaz, H. A., Fernández-Alacid, L. & Pardo-Carrasco, S. C. Biomarkers in skin mucus for a minimally invasive approach to stress in red tilapia (*Oreochromis* sp.) *Fry Biol*. 14, 112. (2025). doi.org/10.3390/biology14020112
34. Beckmann, M. J., Saraiva, M., McLaggan, D., Pottinger, T. G. & van West, P. Saprolegnia infection after vaccination in Atlantic salmon is associated with differential expression of stress and immune genes in the host. *Fish. Shellfish Immun.* **106**, 1095–1105 (2020).
35. Srivastava, V., Rezinciuc, S. & Bulone, V. Quantitative proteomic analysis of four developmental stages of *Saprolegnia parasitica*. *Front. Microbiol.* **8**, 2658. 10.3389/fmicb.2017.02658 (2018).
36. Wang, C. et al. Effects of dietary glutathione on the growth performance, skin mucus antioxidant capacity, and immune responses of juvenile taimen Hucho taimen. *Aquacult. Int.* **30**, 2871–2884. doi.org/10.1007/s10499-022-00939-y (2022).
37. Matthews, E., Ellison, A. & Cable, J. *Saprolegnia parasitica* zoospore activity and host survival indicates isolate variation in host preference. *Fungal Biol.* **125**, 260–268 (2021).
38. Casas-Mulet, R., Matthews, E., Geist, J., Durance, I. & Cable, J. Negative effects of parasite exposure and variable thermal stress on brown trout (*Salmo trutta*) under future climatic and hydropower production scenarios. *Clim. Change Ecol.* **2**, 100039 (2021).
39. Stewart, A. et al. Half the story: Thermal effects on within host infectious disease progression in a warming climate. *Global Change Biol.* **24**, 371–386. doi.org/10.1111/gcb.13842 (2018).
40. Olsen, A. S., Rosin, P. L., Jones, C. B., Cable, J. & Perkins, S. E. Computer vision for infectious disease surveillance; *Saprolegnia* spp. in salmonids. *Ecol. Inf.* **93**, 103567. doi.org/10.1016/j.ecoinf.2025.103567 (2026).
41. Zahran, E. & Risha, E. Protective role of adjuvant and potassium permanganate on oxidative stress response of Nile tilapia (*Oreochromis niloticus*) challenged with *Saprolegnia ferax*. *SpringerPlus* 2, 94. (2013).
42. Noga, E. J. *Fish disease: diagnosis and treatment. Iowa* (Wiley, 2010).

Figures

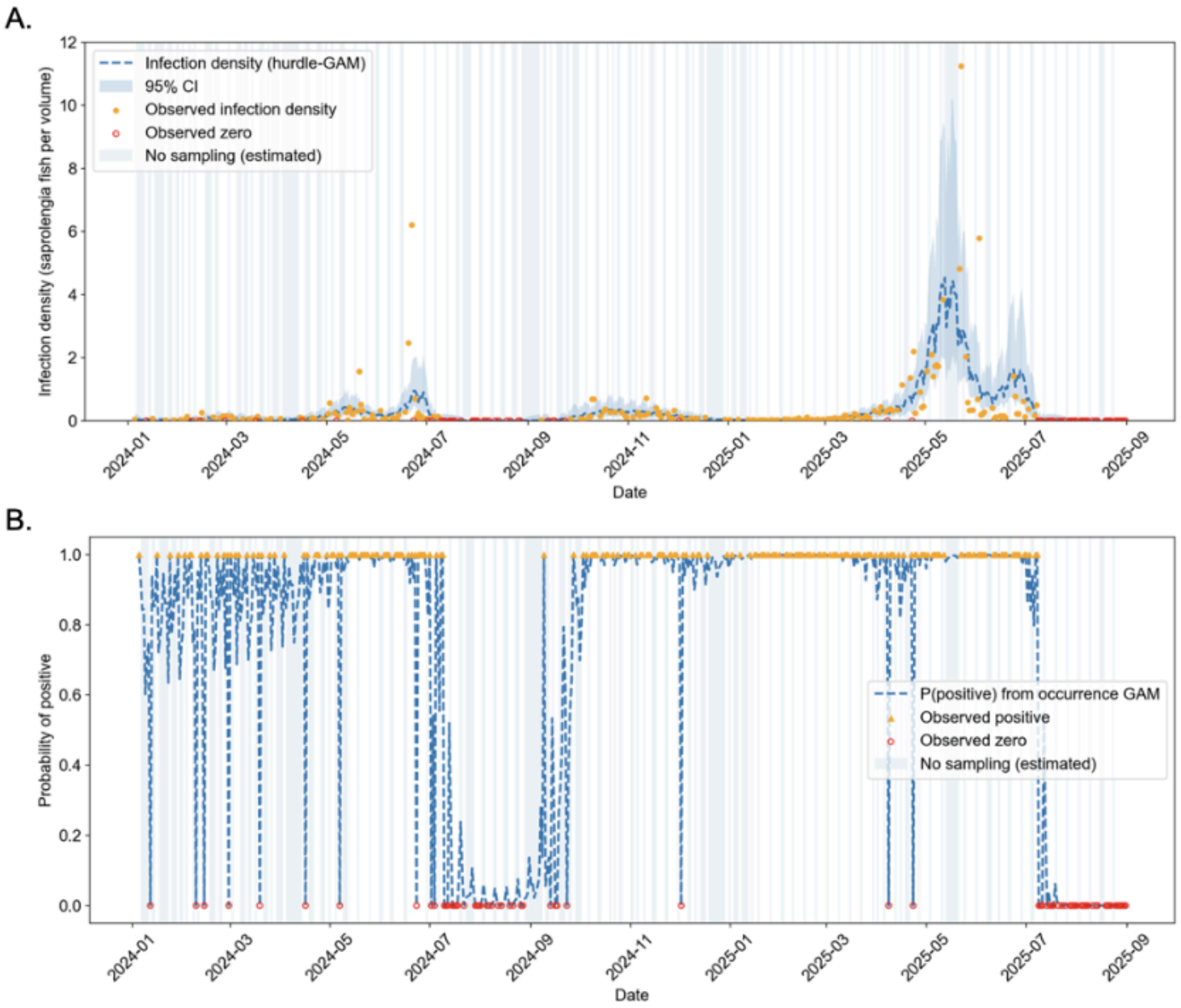


Figure 1

Walk-forward hurdle-GAM for infection estimation. (A) Daily infection density (infected fish per pool volume) estimated using a two-part hurdle Generalised Additive Model (GAM) applied in a walk-forward, causal manner. The dashed blue line indicates the model mean, and the shaded band represents 95 % confidence limits derived from random draws of the fitted coefficients. Orange circles denote observed infection densities on days with sampling effort (Effort = 1), while open red circles indicate observed zeros (Effort = 1). Light grey shaded areas mark periods without sampling (Effort = 0), during which infection densities were model-imputed and may have been constrained to zero by the calibrated limit-of-detection + hysteresis rule. **(B)** Daily probability of infection presence estimated from the logistic (occurrence) component of the hurdle GAM, fitted in the same walk-forward manner. The dashed line shows the predicted probability of a positive observation, orange triangles denote observed positives (Removed > 0 on Effort = 1 days), and open red circles mark observed zeros (Removed = 0 on Effort = 1).

Light grey shaded areas indicate no-sampling periods (Effort = 0), where probabilities were extrapolated solely from the model.

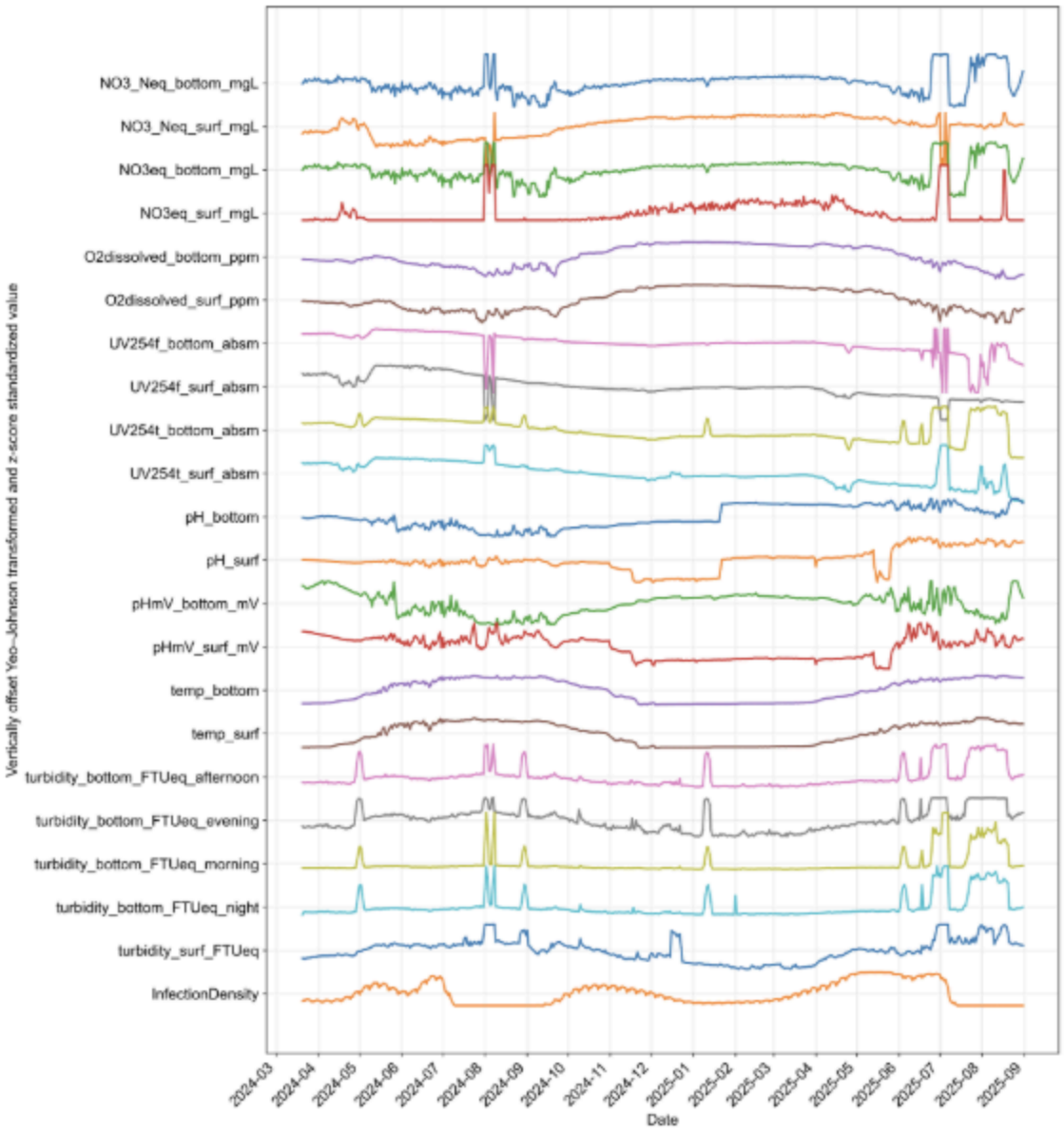


Figure 2

Time series of water-quality features and infection density. All variables were Yeo–Johnson transformed and winsorised at the 1st–99th percentiles to reduce the flattening effect of extreme peaks while preserving relative variation. Each line represents a standardised feature (z-score), with infection density shown last.

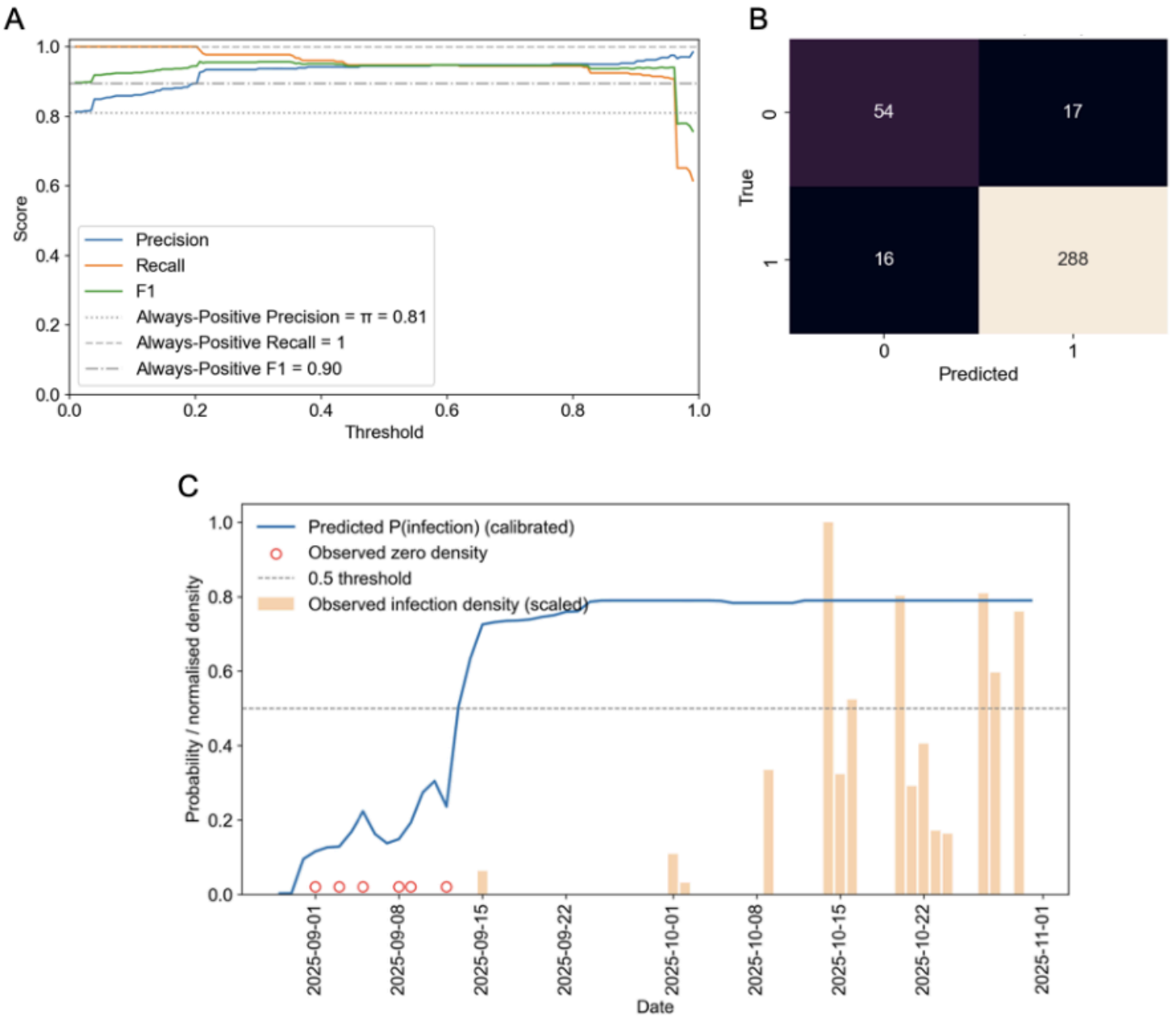


Figure 3

Threshold robustness and early-warning performance of the infection-risk model. (A) Precision, recall, and F1-score as functions of the probability threshold. F1 remains high across a broad threshold range, indicating stable predictive performance and limited sensitivity to the exact threshold choice. The dashed horizontal line denotes the naïve "always-infected" baseline F1. **(B)** Confusion matrix at a fixed probability threshold (t = 0.5), illustrating the balance between correctly identified infection and non-infection days under a conventional midpoint classification rule. **(C)** Infection probability forecasts for an independent test period (September–October 2025). The blue line shows predicted infection probability, orange bars indicate observed infection density (normalised for visual comparison), and open red circles mark days with zero observed infection. The dashed horizontal line denotes the operational decision

threshold (0.5). Infection probability begins to rise approximately 15 days before the first observed infection.

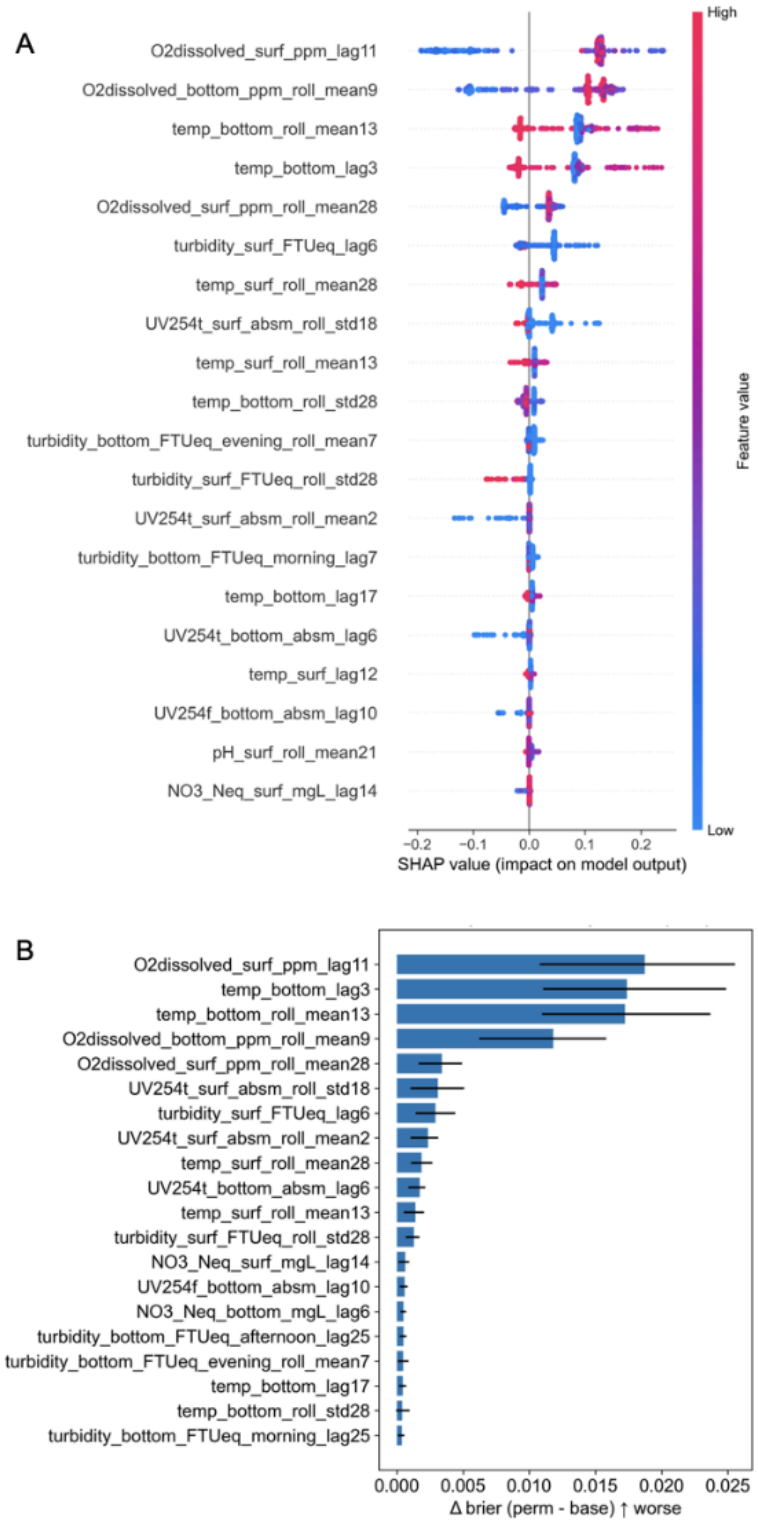


Figure 4

Feature importance and model interpretability of the final Random Forest model. **(A)** SHAP beeswarm plot showing the 20 most influential features ranked by their mean absolute contribution to infection

probability. Each point represents one observation, with colour indicating the relative feature value (blue = low, red = high). Positive SHAP values increase the predicted probability of infection, while negative values decrease it. **(B)** Block permutation importance based on Brier score degradation for the same 20 features. Bars show the mean increase in Brier score when each variable is shuffled within 14-day temporal blocks, with error bars denoting 95% confidence intervals across 200 permutations.

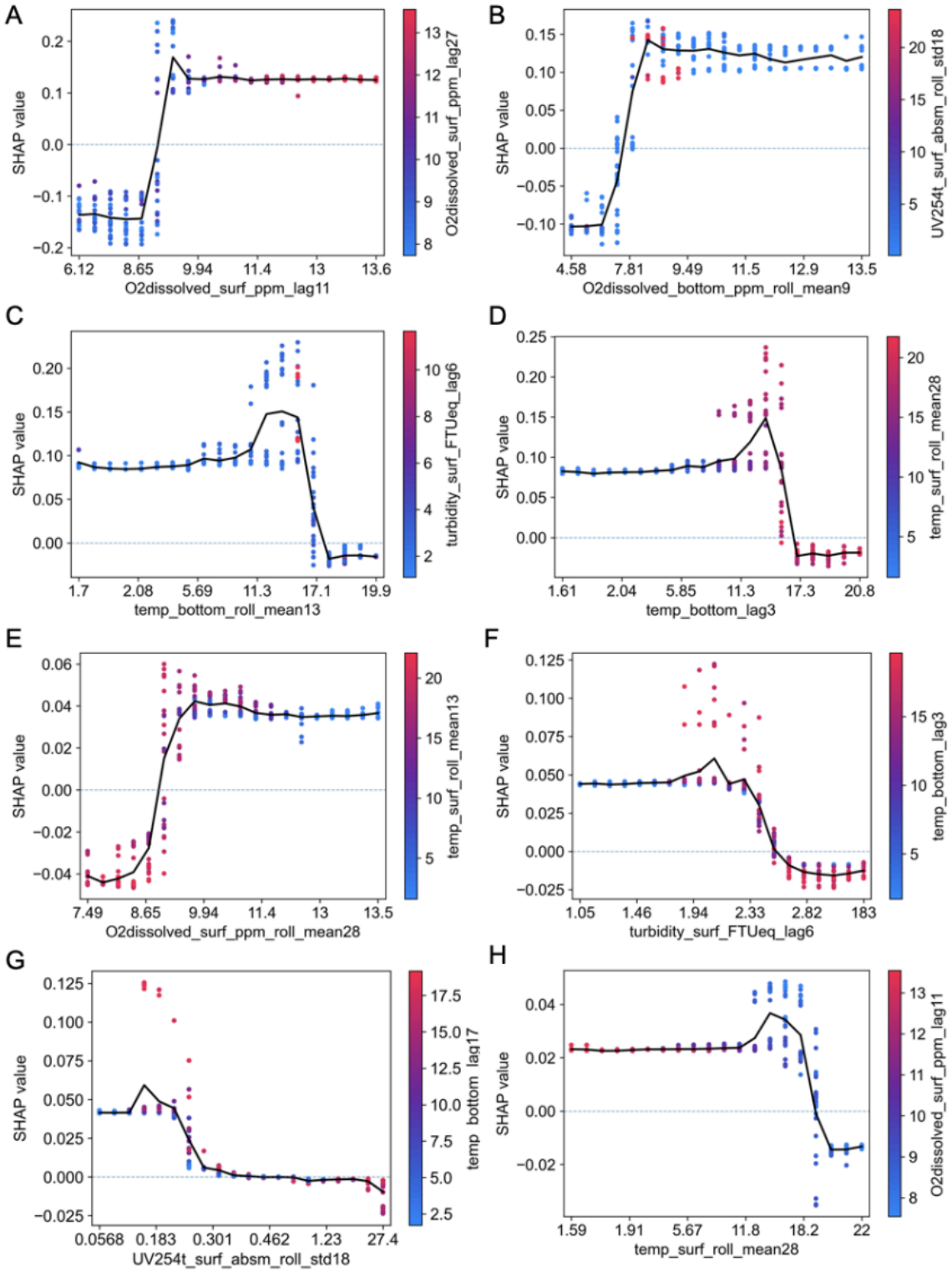


Figure 5

Quantile-based SHAP dependence plots for the 16 most important features. Each point shows a sample's SHAP value (y-axis) plotted against the quantile position of the feature (x-axis), providing even spacing along the horizontal axis and reducing the influence of extreme raw values. The secondary x-axis shows the corresponding raw feature values for easier interpretation. Colours indicate the feature that most strongly interacts with the one on the x-axis, as identified automatically by SHAP. The solid black line represents the mean SHAP value within each quantile, and the dashed horizontal line marks zero contribution to infection probability.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.docx