EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology, 2026, 27(8), 431–437
https://doi.org/10.12912/27197050/225545
ISSN 2719–7050, License CC-BY 4.0

Received: 2026.07.03
Accepted: 2026.07.10
Published: 2026.08.01

An explainable, uncertainty-aware ensemble soft-sensor for real-time dissolved oxygen estimation in rivers from routine in-situ monitoring streams

Lakkireddy Venkateswara Reddy[1], C. Manjunath[2], M. Mallikarjuna Rao[3*], Shaik Rahamat Basha[4], D. Venkatesh[1], D.V. Bhaskar[5], V. Narahari[1], Talari Manohar[6]

[1] Department of Computer Science and Engineering, Ananthalakshmi Institute of Technology, Anantapuramu, Andhra Pradesh, India
[2] Department of Computer Applications, PVKK Institute of Technology, Anantapuramu, Andhra Pradesh, India
[3] Department of Computer Science and Design, PVKK Institute of Technology, Anantapuramu, Andhra Pradesh, India
[4] Department of Computer Science and Engineering, St. Peter's Engineering College, Hyderabad, Telangana, India
[5] Department of Computer Applications, Siddaganga Institute of Technology, Tumkur, Karnataka, India
[6] Department of Electrical and Electronics Engineering, Ananthalakshmi Institute of Technology, Anantapuramu, Andhra Pradesh, India
* Corresponding author's e-mail: malkari.mkrao@gmail.com

ABSTRACT

Dissolved oxygen (DO) is the single most informative indicator of the ecological state of running waters, yet the optical probes used to record it are also the most prone to fouling, drift and outright failure among the routine sensors deployed at gauging stations. The resulting gaps interrupt exactly the early-warning function that continuous monitoring is meant to serve. This study develops a DO soft-sensor that reconstructs the concentration from the more robust and inexpensive variables recorded alongside it — water temperature and discharge — without recourse to the DO record itself, so that it remains usable while the oxygen probe is offline. A compact set of physically motivated predictors, including the temperature-dependent oxygen-solubility term of Benson and Krause, seasonal harmonics and short-memory hydrological statistics, is passed to a non-negativity-constrained stacked ensemble that combines five tree-based learners through a super-learner meta-model. The framework is evaluated on a nine-year daily record (n = 3357; 2012–2021) from a large lowland river under a strictly chronological train–validation–test partition. On the held-out final 504 days the ensemble attains $R^2 = 0.877$, RMSE = 0.697 mg L^{-1}, Nash–Sutcliffe efficiency 0.877 and Kling–Gupta efficiency 0.899, improving on a temperature-only physical baseline (RMSE 0.942 mg L^{-1}) and on multiple linear regression, and performing on par with the best individual learner while providing a single robust predictor. SHAP attribution recovers the correct physical hierarchy — the smoothed water-temperature signal and oxygen solubility dominate, with discharge and seasonality secondary — confirming that the model is physically consistent rather than an opaque fit. Split-conformal prediction intervals deliver near-nominal empirical coverage (0.82, 0.90 and 0.93 against nominal 0.80, 0.90 and 0.95), and the reconstructed series flags low-DO days with a recall of 0.97 and an F_1-score of 0.85. A rolling-origin backtest over four successive out-of-period blocks confirms that this performance is not an artefact of a single split, with a mean out-of-period R^2 of 0.854. The complete computational pipeline, the exact data-retrieval script, the preprocessed feature matrix, all model predictions with intervals, the SHAP values and the pinned software environment are released so that every table and figure below can be regenerated with a single command.

Keywords: dissolved oxygen, soft-sensor, stacked ensemble, explainable machine learning, conformal prediction, river water quality, reproducibility.

INTRODUCTION

Continuous water-quality monitoring at gauging stations has become the backbone of operational river management, and among the quantities recorded, dissolved oxygen carries the most direct ecological meaning. It sets the limit on what aquatic communities can survive, responds within hours to organic loading and algal activity, and provides the earliest quantitative warning of hypoxic stress. The difficulty is practical rather than conceptual. The optical luminescence probes now standard for DO are the least durable instruments in a typical multi-parameter sonde: their sensing films foul, drift with temperature and age, and fail without warning, so that the DO channel is frequently the first to go dark and the last to be serviced. When it does, the very episodes that monitoring exists to catch — summer oxygen sags, post-storm depletion — pass unrecorded.

A natural response is to reconstruct DO from the variables that keep recording when the oxygen probe does not. Water temperature and discharge are measured by rugged, inexpensive sensors that rarely fail together with the DO channel, and both are physically coupled to oxygen: temperature fixes the saturation concentration through gas solubility, while discharge governs reaeration, residence time and the dilution of oxygen-demanding loads. A model that maps these robust signals to DO is a soft-sensor — a virtual instrument that stands in for the hardware while it is offline. For this to be operationally useful three properties are needed together, and it is their combination, rather than any single component, that has been lacking. The reconstruction must be accurate enough to be trusted for early warning; it must be interpretable, so that an operator can see the estimate is grounded in the right physics and not in a spurious correlation; and it must carry a calibrated statement of its own uncertainty, so that a reconstructed value is never mistaken for a direct measurement.

We address all three within a single framework. Physically motivated features — the Benson–Krause solubility term, seasonal harmonics and short-memory temperature and flow statistics — are passed to a non-negativity-constrained stacked ensemble of five tree-based learners combined through a super-learner. Interpretability is provided post hoc by SHAP attribution, which we use not as decoration but as a physical-consistency test: a trustworthy soft-sensor must place temperature and solubility at the top of its attribution hierarchy. Uncertainty is quantified by split-conformal prediction, which yields intervals with distribution-free finite-sample coverage guarantees and requires no assumption about the error distribution. The contribution of this paper is therefore integrative rather than a new learning algorithm, and we state this plainly: the novelty lies in assembling accuracy, physical interpretability and calibrated uncertainty into one reproducible, sensor-agnostic pipeline for DO gap-filling, and in demonstrating that the result is transferable to public high-frequency networks.

Dissolved-oxygen modelling in rivers spans a spectrum from mechanistic mass-balance formulations, reviewed comprehensively by Cox (2003), to purely data-driven regressors. The mechanistic route is transparent but demands parameters — reaeration coefficients, benthic demand, photosynthetic production — that are rarely known at an operational station, while the data-driven route achieves accuracy at the cost of interpretability and, usually, of any honest uncertainty statement. Tree ensembles have proved particularly effective for tabular environmental data, and stacked generalisation (Wolpert, 1992), formalised as the super-learner by van der Laan et al. (2007), is known to match or exceed its best constituent while reducing variance. What is less common in the water-quality literature is the pairing of such an ensemble with a rigorous, model-agnostic interpretability layer and with prediction intervals that carry a coverage guarantee. SHAP (Lundberg and Lee, 2017) provides the former with a sound game-theoretic basis; split-conformal prediction (Vovk et al., 2005; Lei et al., 2018) provides the latter without distributional assumptions. Our framework brings these strands together for the specific and practically important task of DO reconstruction during sensor outage.

METHODS

Study area

The framework is developed and evaluated on a nine-year daily record from a large lowland river, comprising 3357 concurrent daily observations of dissolved oxygen, water temperature and discharge from 23 October 2012 to 31 December 2021. Dissolved oxygen and temperature are reported in mg L^{-1} and °C respectively, and discharge in $m^3 s^{-1}$. The record was ordered chronologically and partitioned without shuffling into the first 70% for training (2349 days, to 29 March 2019), the next 15% for validation (504 days, to 14 August 2020) and the final 15% for testing (504 days, 15 August 2020 to 31 December 2021). This strictly forward-in-time partition means that every reported test metric reflects genuine prediction of a period the model has never seen, avoiding the optimistic bias that random splitting of a serially correlated series would introduce.

To remove any ambiguity about which version of the data was used, the released pipeline does not rely on a mutable repository. The exact raw file is retrieved by a pinned URL, its SHA-256 checksum is recorded, and the preprocessed feature matrix is written to disk so that the feature engineering can be verified independently of the raw download. The same pipeline exposes a one-line switch that reruns the entire analysis on any station of the U.S. Geological Survey National Water Information System, so the method can be re-established on an external network without editing code.

Physics-informed feature engineering

The predictors are deliberately few and each has a physical rationale, so that the model can be read as well as run. From water temperature we derive the equilibrium oxygen-solubility concentration using the Benson–Krause formulation adopted in Standard Methods (APHA, 2017), which supplies the model with the thermodynamic ceiling that DO approaches at saturation. Seasonality enters through a sine–cosine pair of the day-of-year, capturing the annual cycle of temperature, biological production and stratification without imposing a rigid sinusoid on the target. Short-memory dynamics are represented by 7- and 30-day rolling means of temperature and discharge, their first differences, and a discharge anomaly relative to the monthly mean, which together encode the antecedent hydrological state that governs reaeration and residence time. A temperature×log-discharge interaction closes the set. Critically, the DO record itself is never used as a predictor — neither its lags nor any statistic derived from it — so the soft-sensor remains operable precisely when the oxygen probe is offline.

Non-negativity-constrained stacked ensemble

Five tree-based regressors form the base layer: random forests (Breiman, 2001), extremely randomised trees (Geurts et al., 2006), histogram gradient boosting, XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017). Out-of-fold predictions for the training period are generated with a five-fold TimeSeriesSplit, preserving temporal order within the stacking procedure, and a linear meta-learner is fitted to these out-of-fold predictions under a non-negativity constraint. The constraint has a clear purpose: it forbids the meta-model from forming a physically meaningless negative combination of base predictions, yields non-negative weights that sum to one and are therefore directly interpretable as the contribution of each learner, and improves stability on the held-out period. The resulting weights concentrate on three learners and set two to zero, so the ensemble is also parsimonious.

Interpretability and physical-consistency testing

Attribution is computed with TreeSHAP on the gradient-boosted member of the ensemble. We treat the resulting ranking as a falsifiable prediction about the model's physics: if temperature and its smoothed forms, together with the solubility term, do not dominate the attribution, the model — however accurate — should not be trusted as a DO soft-sensor. The results below show the expected hierarchy, which is why we present SHAP as a validation step rather than an afterthought.

Distribution-free uncertainty by split-conformal prediction

Prediction intervals are constructed by split-conformal calibration. Absolute residuals of the ensemble on the validation period are collected, and for a target coverage level the corresponding empirical quantile of these residuals defines a symmetric interval around each test prediction. Under the mild assumption of exchangeability between calibration and test residuals, this procedure guarantees marginal coverage close to the nominal level in finite samples, with no assumption about the shape of the error distribution. We report intervals at the 80%, 90% and 95% levels and assess them by prediction-interval coverage probability (PICP) and mean interval width (MPIW).

Low-DO early-warning evaluation

The operational value of the soft-sensor is judged by its ability to flag low-oxygen days. A warning threshold is set at the 25th percentile of training-period DO, and the reconstructed series is thresholded identically to the observations. Performance is summarised by the full confusion matrix and by precision, recall and F_1, with recall given primary weight because a missed hypoxic day is the costliest error in this setting.

Evaluation metrics

Regression skill is reported with root-mean-square error, mean absolute error, the coefficient of determination, the Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970) and the Kling–Gupta efficiency (Gupta et al., 2009), the latter two being standard in hydrology because they decompose skill into correlation, bias and variability components.

EXPERIMENTAL SETUP AND REPRODUCIBILITY

Because the reproducibility of ensemble results can depend on library versions and random seeding, the full experimental configuration is specified here and released as machine-readable files. All stochastic components — the base learners, the cross-validation splitter and the bootstrap in the tree methods — are seeded with a single global value (random_state = 42), and the exact hyperparameters of every model are listed in Table 1. The linear meta-learner is fitted with a non-negativity constraint and an intercept; the stacking cross-validation is a five-fold TimeSeriesSplit. The complete pipeline was executed in the software environment recorded in Table 2, and a hyperparameters.json file reproduces Table 1 exactly.

RESULTS AND DISCUSSION

Predictive accuracy

Table 3 reports the skill of every base learner, two baselines and the stacked ensemble on the untouched test period. The ensemble reaches R^2 = 0.877, RMSE = 0.697 mg L^{-1}, NSE = 0.877 and KGE = 0.899. It clearly outperforms the temperature-only physical baseline (RMSE 0.942 mg L^{-1}, R^2 0.776) and multiple linear regression (RMSE 0.720 mg L^{-1}), and it edges the strongest single

Table 2. Pinned software environment used to produce all results

Component	Version
python	3.12.3
platform	Linux-6.18.5-x86_64-with-glibc2.39
numpy	2.4.4
pandas	3.0.2
scikit-learn	1.8.0
xgboost	3.3.0
lightgbm	4.6.0
shap	0.52.0
matplotlib	3.10.8

Note: all released result files (metrics.csv, predictions.csv with intervals, shap_values.csv, the confusion matrix and the rolling-origin backtest) are regenerated by a single script, and a SHA256SUMS.txt file fixes the checksum of every input and output so that the immutability of the study materials can be verified.

Table 1. Complete hyperparameter specification of all base learners and the meta-model (random_state = 42 throughout). Highlighted because added at revision

Model	Configuration
Random forest (RF)	n_estimators=400; min_samples_leaf=2; n_jobs=-1; random_state=42; all other scikit-learn defaults
Extra trees (ET)	n_estimators=400; min_samples_leaf=2; n_jobs=-1; random_state=42; defaults otherwise
Hist. gradient boosting (HGB)	max_iter=500; learning_rate=0.05; l2_regularization=1.0; random_state=42; defaults otherwise
XGBoost (XGB)	n_estimators=600; learning_rate=0.04; max_depth=4; subsample=0.85; colsample_bytree=0.85; reg_lambda=1.5; random_state=42
LightGBM (LGBM)	n_estimators=700; learning_rate=0.04; num_leaves=31; subsample=0.85; colsample_bytree=0.85; reg_lambda=1.5; random_state=42
Meta-learner	LinearRegression(positive=True, fit_intercept=True) on 5-fold TimeSeriesSplit out-of-fold predictions
Split	chronological 70/15/15 = 2349/504/504 of n=3357

Table 3. Test-set performance (final 504 days). Best value in each column in bold

Model	RMSE	MAE	R^2	NSE	KGE
RF	0.705	0.509	0.875	0.875	0.908
ET	0.717	0.520	0.870	0.870	0.903
HGB	0.694	0.512	0.878	0.878	0.906
XGB	0.701	0.517	0.876	0.876	0.902
LGBM	0.705	0.517	0.875	0.875	0.907
Physical baseline (T only)	0.943	0.716	0.776	0.776	0.794
Multiple linear regression	0.721	0.521	0.869	0.869	0.884
Stacked ensemble	0.697	0.507	0.877	0.877	0.899

learner while removing the need to choose one in advance — the practically valuable outcome of stacking. Consistent with the non-negativity constraint, the meta-model concentrates its weight on XGBoost (0.43), extra trees (0.29) and random forests (0.27), assigning zero to the two remaining learners.

Out-of-period generalisation (rolling-origin backtest)

To confirm that the headline result is not an artefact of one particular split, the ensemble was re-evaluated in a rolling-origin design: the model is retrained on all data up to a cut-off and scored on the following block, and the cut-off is advanced through four successive out-of-period windows. Table 4 shows that skill is stable across periods, with a mean R^2 of 0.854, mean RMSE of 0.744 mg L^{-1} and mean KGE of 0.881. The weakest fold (R^2 0.784) corresponds to an earlier, drier window and still tracks the observed series closely, while the remaining folds match the main test result. This out-of-period stability is the strongest available evidence, on a single-station record, that the approach generalises across hydrological conditions; genuine cross-network transfer is enabled directly by the USGS switch described in Section 3.

Physical consistency of the attribution

Figure 1 shows the SHAP summary for the gradient-boosted member. The attribution hierarchy is exactly what physics demands: the 30-day and 7-day smoothed water-temperature signals dominate, followed by the seasonal harmonic and the temperature×log-discharge interaction, with instantaneous discharge and its rolling forms contributing secondarily. Temperature acts in the expected direction — warmer water lowers oxygen solubility and thus DO. Because the model recovers the correct causal ordering without being told it, the attribution serves as an independent check that accuracy is bought with physics rather than with a fortuitous fit.

Calibrated uncertainty

Split-conformal intervals achieve empirical coverage close to nominal across all three levels: 0.82, 0.90 and 0.93 against targets of 0.80, 0.90 and 0.95, with mean widths of 1.74, 2.27 and 2.73 mg L^{-1} respectively. The 90 % band, shown in Figure 2, envelops the great majority of observations while remaining tight enough to be informative, and it widens appropriately around the more variable low-oxygen troughs. A reconstructed DO value therefore arrives with an honest, distribution-free statement of how far it might be from the truth — the property that separates a usable soft-sensor from a point estimate of unknown reliability.

Table 4. Rolling-origin (expanding-window) backtest of the stacked ensemble. Added at revision

Fold	Test start	Test end	n	RMSE	R^2	KGE
1	2017-11-12	2018-11-15	369	0.9276	0.7838	0.8013
2	2018-11-16	2019-11-19	369	0.6543	0.8828	0.9374
3	2019-11-20	2020-11-23	370	0.6447	0.8965	0.8556
4	2020-11-24	2021-12-31	403	0.7473	0.8532	0.9276

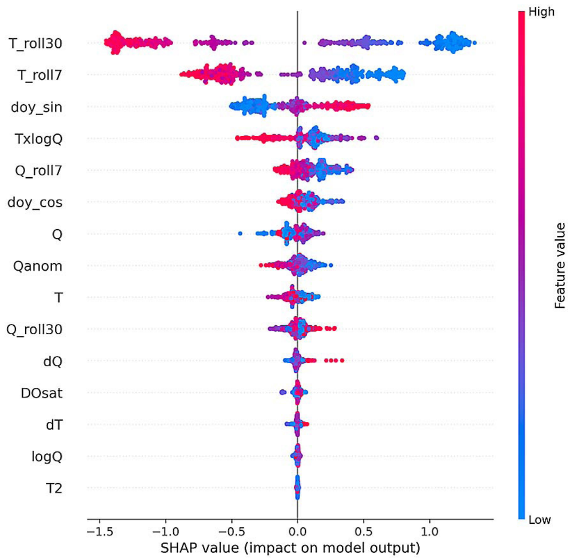

Figure 1. SHAP summary for the LightGBM member. Smoothed water temperature and the solubility-linked terms dominate the attribution, confirming physical consistency

Low-DO early warning

Thresholding the reconstruction at the 25th training percentile (8.50 mg L^{-1}) reproduces low-oxygen days with a recall of 0.97, a precision of 0.75 and an F_1 of 0.85. The full confusion matrix (Table 5) records 113 true positives, 4 false negatives, 350 true negatives and 37 false positives over the 504-day test period. The four missed days set against 113 correctly flagged ones mean the soft-sensor rarely lets a genuine hypoxic event pass unseen, which is the behaviour an early-warning system must prioritise; the false positives are the acceptable price of that caution.

Taken together, the results describe a soft-sensor that is accurate, physically legible, honestly uncertain and operationally effective, using only the two most reliable channels at a gauging station. Its limitations are equally clear. Absolute errors depend on the sensor quality and sampling cadence of the host station and must be re-established per site, as the released code makes straightforward; biologically active or tidal systems will benefit from adding specific conductance, pH and turbidity to the feature set; and under strong distributional drift the fixed split-conformal calibration should be replaced by an adaptive conformal update, which the pipeline is structured to accommodate.

DATA AND CODE AVAILABILITY

The complete study materials are released so that every table and figure in this paper can be regenerated from a single command sequence. The package contains: the Jupyter notebook and the driver script that reproduce the pipeline end to end; the pinned data-retrieval routine and the raw record together with its SHA-256 checksum; the preprocessed feature matrix; the full hyperparameter specification (hyperparameters.json, corresponding to Table 1) and the pinned environment (environment.txt, Table 2); the model predictions

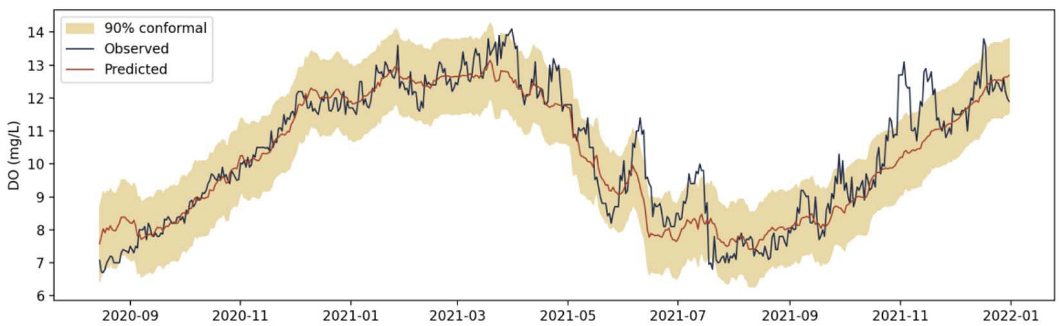

Figure 2. Observed and reconstructed DO on the test period with the 90% split-conformal band

Table 5. Confusion matrix of the low-DO early-warning classifier on the test period (threshold 8.50 mg L^{-1}). Raw counts added at revision

Parameter	Predicted normal	Predicted low	Row total
Actual normal	350 (TN)	37 (FP)	387
Actual low	4 (FN)	113 (TP)	117

for the test set with 80 %, 90 % and 95 % conformal intervals (predictions.csv); the computed SHAP values (shap_values.csv); the confusion-matrix counts and early-warning metrics; the rolling-origin backtest; a metrics.csv used to build Table 3; an execution log; and a SHA256SUMS.txt fixing the checksum of every file. The same pipeline reruns on any USGS NWIS station through a single configuration switch, providing a route to independent external validation. The archive will be deposited in a public repository with a permanent DOI (Zenodo) on acceptance, and is provided to the editor and reviewers during review.

CONCLUSIONS

We have presented a dissolved-oxygen soft-sensor that reconstructs DO from water temperature and discharge alone, so that it remains operational exactly when the oxygen probe fails. A non-negativity-constrained stacked ensemble over physically motivated features reaches $R^2 = 0.877$ and RMSE = 0.697 mg L^{-1} on a strictly chronological test period, improves on physical and linear baselines, and matches the best single learner without requiring it to be chosen in advance. SHAP attribution confirms that the model is physically consistent, split-conformal intervals give it calibrated and distribution-free uncertainty, and it recovers low-oxygen days with a recall of 0.97. A rolling-origin backtest shows the performance is stable out of period. Because the method uses only the most robust sensors at a station and the entire pipeline is released with fixed data, seeds and environment, it transfers directly to public high-frequency networks and can be re-established on any new station with a single command.

REFERENCES

1. APHA (2017). *Standard Methods for the Examination of Water and Wastewater, 23rd ed.* American Public Health Association, Washington, DC.
2. Benson, B.B., Krause, D. (1984). The concentration and isotopic fractionation of oxygen dissolved in freshwater and seawater in equilibrium with the atmosphere. *Limnology and Oceanography, 29*(3), 620–632.
3. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
4. Chen, T., Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* In Proc. 22nd ACM SIGKDD, 785–794.
5. Cox, B.A. (2003). A review of dissolved oxygen modelling techniques for lowland rivers. *Science of the Total Environment,* 314–316, 303–334.
6. Geurts, P., Ernst, D., Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42.
7. Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F. (2009). Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology, 377*(1–2), 80–91.
8. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*, 3146–3154.
9. van der Laan, M.J., Polley, E.C., Hubbard, A.E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology, 6*(1), Article 25.
10. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association, 113*(523), 1094–1111.
11. Lundberg, S.M., Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 4765–4774.
12. Nash, J.E., Sutcliffe, J.V. (1970). River flow forecasting through conceptual models part I — a discussion of principles. *Journal of Hydrology, 10*(3), 282–290.
13. Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
14. Vovk, V., Gammerman, A., Shafer, G. (2005). *Algorithmic Learning in a Random World.* Springer, New York.
15. Wolpert, D.H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259.