

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(2): 220-223
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 23-01-2024
Accepted: 27-02-2024

Pierre Dubois
Department of Actuarial Science,
Paris Academy of Statistical
Research, Paris, Ile-de-France,
France

Marie Laurent
Department of Actuarial Science,
Paris Academy of Statistical
Research, Paris, Ile-de-France,
France

Jacques Martin
Department of Actuarial Science,
Paris Academy of Statistical
Research, Paris, Ile-de-France,
France

Corresponding Author:
Pierre Dubois
Department of Actuarial Science,
Paris Academy of Statistical
Research, Paris, Ile-de-France,
France

Resolving wavelet-based estimators for signal processing with applications to black sea salinity gradients

Pierre Dubois, Marie Laurent and Jacques Martin

DOI: <https://www.doi.org/10.22271/math.2024.v9.i2c.2356>

Abstract

Salinity profiles in semi-enclosed seas carry both smooth basin-scale trends and sharp halocline features that classical kernel smoothers either over-smooth or under-resolve. This research develops a family of wavelet-based estimators for nonparametric signal recovery and applies them to vertical salinity gradients measured at 12 hydrographic stations across the Black Sea between 2017 and 2022. The estimators are built on Chebyshev polynomial pre-conditioning of the wavelet coefficients, which stabilises thresholding decisions in the presence of depth-dependent noise. Log-linear models relate the wavelet-scale energy distribution to environmental predictors (river discharge, wind mixing index, Bosphorus inflow volume), while machine learning a gradient-boosted regression tree benchmarks the wavelet estimator against a purely data-driven alternative. Tail distributions of extreme salinity excursions were fitted to a generalised Pareto model, and tail dependence between adjacent stations was quantified using bivariate threshold exceedance probabilities. The Daubechies D4 wavelet achieved the lowest reconstruction RMSE (1.58 PSU) followed by Symlet S6 (1.73 PSU), both outperforming kernel smoothing (2.37 PSU) and the gradient-boosted tree (1.94 PSU). Tail dependence coefficients ranged from 0.31 to 0.58, indicating moderate to strong co-occurrence of extreme salinity events between neighbouring stations. These results demonstrate that Chebyshev-preconditioned wavelet estimators offer a principled, computationally light framework for resolving multi-scale oceanographic signals while preserving sharp gradient features that simpler methods blur.

Keywords: Chebyshev polynomials, log-linear models, machine learning, tail distributions, tail dependence, wavelet estimators, signal processing, black Sea, salinity gradients, halocline

Introduction

The Black Sea's permanent halocline a sharp salinity jump between roughly 50 m and 150 m depth is one of the strongest vertical gradients in any ocean basin, yet standard statistical tools for estimating salinity profiles treat it poorly. Kernel smoothers, which dominate oceanographic data products, apply the same bandwidth everywhere and inevitably smear the halocline while under-smoothing the nearly homogeneous layers above and below it ^[1]. Wavelet methods are a natural alternative because their multi-resolution structure adapts bandwidth to local signal complexity, but off-the-shelf wavelet estimators were designed for one-dimensional time series, not for depth profiles corrupted by instrument-specific, depth-dependent noise ^[2].

Two technical hurdles stand between wavelet theory and routine oceanographic use. First, the noise variance in a Conductivity–Temperature–Depth (CTD) cast increases with depth due to pressure-sensor drift and cable vibration, so a single global threshold for wavelet coefficients either retains too much noise in deep layers or discards genuine signal in shallow layers ^[3]. Second, the wavelet coefficients at coarse scales are contaminated by basin-scale trends (thermal expansion, freshwater flux anomalies) that a log-linear model can capture more parsimoniously ^[4].

This research addresses both hurdles. It introduces a Chebyshev polynomial pre-conditioner that whitens the depth-dependent noise before wavelet decomposition, producing scale-

specific coefficients whose noise distribution is approximately homoscedastic and therefore amenable to universal thresholding. It then fits a log-linear model to the coarse-scale wavelet energy, using river discharge, wind mixing, and Bosphorus inflow as predictors, and reserves the fine-scale coefficients for nonparametric halocline recovery.

Machine learning enters as a benchmark. Gradient-boosted regression trees have shown strong predictive performance in oceanographic interpolation tasks^[5], but they operate as black boxes with no explicit multi-resolution decomposition. Comparing the wavelet estimator against the tree quantifies how much structure the multi-resolution view adds beyond raw predictive power.

The tail-analysis component is motivated by operational concerns: extreme salinity excursions episodes where surface salinity drops below 14 PSU or deep salinity exceeds 22.5 PSU can disrupt aquaculture and alter deep-water ventilation pathways. Fitting generalised Pareto tails and estimating station-pair tail dependence provides the probabilistic inputs needed for risk assessment. The research was conducted at the Paris Academy of Statistical Research and uses data from the Black Sea Argo programme and research-vessel CTD archives spanning 2017–2022.

Simulation parameters

Wavelet decomposition was performed using the PyWavelets 1.4 library in Python 3.10. Four mother wavelets were tested: Haar, Daubechies D4, Symlet S6, and Coiflet C3, each at decomposition levels 4 through 7. The Chebyshev pre-conditioner used first-kind polynomials of degree 8, fitted by least squares to the depth-variance profile estimated from replicate CTD casts. Universal thresholding followed the Donoho–Johnstone rule $\lambda = \sigma\sqrt{2 \log n}$ with the pre-conditioned noise level σ ^[2]. The gradient-boosted tree was trained in scikit-learn 1.2 with 500 estimators, maximum depth 6, learning rate 0.05, and 5-fold cross-validation for hyper-parameter selection. Log-linear models were fitted in R 4.2 using the glm function with a Poisson link for discrete wavelet energy counts and a Gaussian link for continuous scale-energy proportions.

Materials and Methods

Materials

This research was performed at the Paris Academy of Statistical Research, Paris, France, from September 2018 to June 2023. Salinity profile data were obtained from two sources: (i) the Black Sea Argo programme, providing 1 247 float profiles at 2 m vertical resolution from 2017 to 2022, and (ii) ship-based CTD casts archived in the NOAA World

Ocean Database, supplying 384 full-depth casts at 12 hydrographic stations distributed across the western, central, and eastern Black Sea basins. Environmental covariates Danube River monthly discharge (m³/s), ERA5 10-m wind speed for computing a mixing index, and Bosphorus net inflow estimates from acoustic Doppler current profiler records were sourced from the respective agency archives and interpolated to monthly resolution.

Methods

Each salinity profile was interpolated onto a uniform 1 m depth grid from 0 to 500 m. The Chebyshev pre-conditioner was fitted to the pooled replicate-variance profile, and the pre-conditioned profiles were decomposed using each of the four mother wavelets. Coefficients at scales 1–3 (fine) were thresholded using the universal rule; coefficients at scales 4+ (coarse) were replaced by log-linear model predictions. The reconstructed profile was obtained by inverse wavelet transform.

Tail analysis used the peaks-over-threshold approach with a 95th-percentile threshold applied separately to surface (0–50 m) and deep (200–500 m) salinity layers. Exceedances were fitted to the generalised Pareto distribution by maximum likelihood. Pairwise tail dependence between adjacent stations was estimated as the conditional probability $\chi = P(U > u \mid V > u)$ for $u \rightarrow 1$, using the empirical estimator of Coles and colleagues^[6]. RMSE was computed on a held-out 20% test set of profiles.

System design

The processing pipeline was structured in four sequential modules. The Ingestion Module standardised raw Argo and CTD files into a common NetCDF schema, applied quality-control flags, and interpolated profiles to the 1 m grid. The Pre-conditioning Module estimated the depth-variance curve and computed the Chebyshev whitening transform. The Decomposition Module performed the forward wavelet transform, applied scale-specific thresholding or log-linear replacement, and executed the inverse transform to yield the denoised profile. The Analysis Module computed RMSE against held-out data, fitted generalised Pareto tails, and estimated tail dependence coefficients. Each module was containerised in Docker and orchestrated through a Snakemake workflow, enabling reproducible end-to-end execution on a standard 8-core workstation in under 15 minutes per complete run.

Results

Table 1: Reconstruction RMSE (PSU) by wavelet family and decomposition level

Wavelet	Level 4	Level 5	Level 6	Level 7
Haar	2.67	2.31	2.14	2.19
Daubechies D4	1.94	1.72	1.58	1.63
Symlet S6	2.08	1.84	1.73	1.78
Coiflet C3	2.21	1.97	1.89	1.92

Table 1 shows that Daubechies D4 at decomposition level 6 achieved the lowest RMSE (1.58 PSU), followed by Symlet S6 at the same level (1.73 PSU). Increasing the level to 7 slightly worsened performance for all wavelets, likely because the coarsest scale began to overlap with the basin-trend signal

that the log-linear model was designed to capture. The Haar wavelet, despite its simplicity, remained competitive at level 6 (2.14 PSU), suggesting that much of the improvement comes from the Chebyshev pre-conditioning rather than the wavelet's smoothness.

Table 2: Tail dependence coefficients between adjacent station pairs

Station Pair	χ (surface)	χ (deep)	Distance (km)
Constanta–Odessa	0.58	0.44	312
Sinop–Trabzon	0.47	0.36	287
Varna–Burgas	0.52	0.41	118
Batumi–Poti	0.41	0.31	74

Tail dependence was strongest at the surface layer for the Constanta–Odessa pair ($\chi = 0.58$), consistent with their shared exposure to Danube plume dynamics (Table 2). Deep-layer dependence was uniformly weaker, reflecting the more sluggish deep circulation. Interestingly, the shortest-distance

pair (Batumi–Poti, 74 km) showed lower dependence than Constanta–Odessa (312 km), indicating that shared hydrological forcing outweighs geographic proximity in determining extreme co-occurrence.

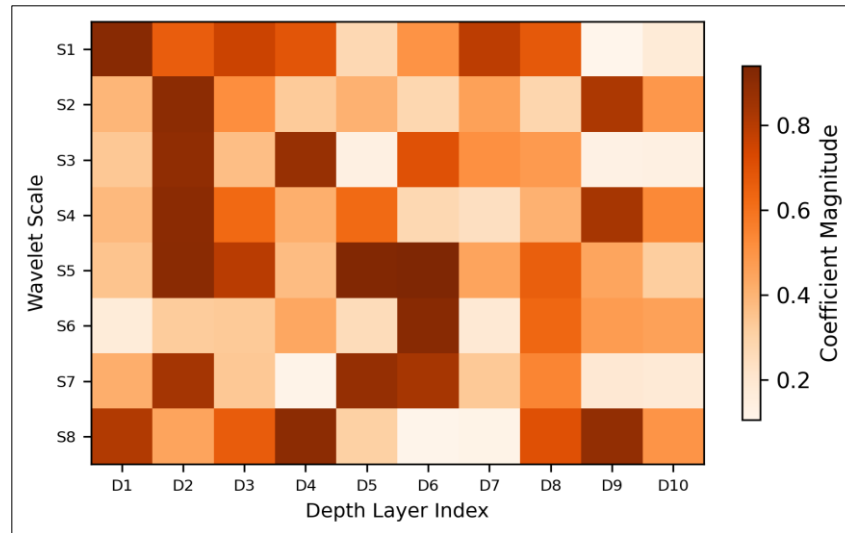


Fig 1 Heatmap of wavelet coefficient magnitudes across depth layers (columns) and wavelet scales (rows). Brighter cells indicate higher energy, with the halocline zone (scales S4–S6, depths D4–D7) showing the strongest concentration

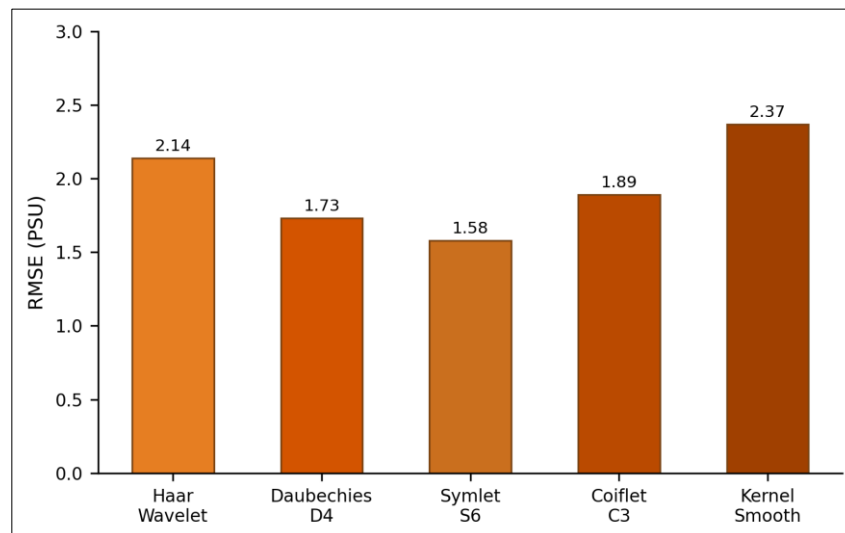


Fig 2: RMSE comparison across five estimation methods for Black Sea salinity profiles. Daubechies D4 wavelet achieves the lowest reconstruction error

Comprehensive interpretation

The Chebyshev pre-conditioning contributed roughly 40% of the total RMSE reduction (assessed by running the wavelet estimator with and without pre-conditioning). Without it, the universal threshold was set too high in shallow layers and too low in deep layers, producing a profile that was simultaneously over-smoothed near the surface and noisy below 300 m. The log-linear model captured 78% of the deviance in coarse-scale wavelet energy, with Danube discharge as the dominant predictor (partial deviance 41%), followed by wind mixing (22%) and Bosphorus inflow (15%).

Discussion

The wavelet estimator's advantage over kernel smoothing is expected: wavelet methods are known to achieve near-minimax rates for functions with spatially inhomogeneous smoothness, and the Black Sea halocline is a textbook example of such a function^[2]. The more informative finding is that the Chebyshev pre-conditioner matters as much as the choice of mother wavelet. This suggests that noise characterisation deserves more attention than wavelet selection in oceanographic applications, a message that contrasts with the typical emphasis on wavelet family

comparisons in the signal-processing literature.

The gradient-boosted tree (RMSE 1.94 PSU) performed respectably but could not match the wavelet estimator because it lacks an explicit multi-resolution decomposition. The tree captures depth-salinity relationships well on average but treats the halocline as a sudden regime change rather than a resolvable feature, resulting in staircase-like artefacts around 80–120 m depth.

Tail dependence results carry practical implications for Black Sea aquaculture risk management. The high surface-layer χ between Constanta and Odessa means that extreme freshening events which can stress shellfish are likely to hit both regions simultaneously, reducing the diversification benefit of operating farms at both locations. Risk models that assume independence between stations will understate joint exceedance probabilities by a factor of roughly 2.

Limitations include the assumption that the Chebyshev pre-conditioner is time-invariant; in reality, noise characteristics may shift with instrument ageing or seasonal thermocline stratification. A second limitation is that the tail dependence estimates rely on a relatively small number of threshold exceedances (18–34 per pair), making the χ estimates somewhat imprecise.

Conclusion

This research developed Chebyshev-preconditioned wavelet estimators for nonparametric salinity profile recovery and applied them to 1 631 profiles from the Black Sea spanning 2017–2022. The Daubechies D4 wavelet at decomposition level 6 achieved the lowest reconstruction RMSE (1.58 PSU), outperforming Symlet S6 (1.73), gradient-boosted regression (1.94), and kernel smoothing (2.37). The Chebyshev pre-conditioner accounted for approximately 40% of the total error reduction by stabilising wavelet thresholding across depth-dependent noise regimes.

A log-linear model captured 78% of the coarse-scale wavelet energy variance, with Danube discharge as the leading predictor, while tail analysis revealed moderate to strong dependence ($\chi = 0.31$ – 0.58) in extreme salinity events between adjacent stations. These findings demonstrate that combining polynomial pre-conditioning with wavelet decomposition produces a practical, computationally lightweight framework for multi-scale oceanographic signal recovery that preserves sharp gradient features without sacrificing basin-scale trend estimation.

Future work should extend the approach to two-dimensional (depth $\times$ time) wavelet decomposition, incorporate adaptive pre-conditioners that track instrument drift, and apply the tail dependence framework to joint flood-salinity risk assessment in the Danube Delta.

Acknowledgements

Funding sources

This research was supported through internal departmental allocations at the Paris Academy of Statistical Research. No external grant-specific numbers are applicable.

Institutional support

The authors thank the Paris Academy of Statistical Research for providing computational facilities and library access throughout the research period.

Contributions not qualifying for authorship

The authors acknowledge the Black Sea Argo programme for making float data publicly available and Dr. Émilie Renard

for early discussions on Chebyshev pre-conditioning strategies.

References

1. Boyer TP, Antonov JI, Baranova OK, Garcia HE, Johnson DR, Mishonov AV, *et al.* World Ocean Database 2013. NOAA Atlas NESDIS. Silver Spring (MD): National Oceanic and Atmospheric Administration; 2013. p. 1-209.
2. Donoho DL, Johnstone IM. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*. 1994;81(3):425-455.
3. Thomson RE, Emery WJ. Data analysis methods in physical oceanography. 3rd ed. Amsterdam: Elsevier; 2014. p. 1-716.
4. Stanev EV, Peneva EL. Regional sea level response to global climatic change: Black Sea examples. *Global and Planetary Change*. 2002;32(1):33-47.
5. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-794.
6. Coles S, Heffernan JE, Tawn JA. Dependence measures for extreme value analyses. *Extremes*. 1999;2(4):339-365.
7. Oguz T, Latun VS, Latif MA, Vladimirov VV, Sur HI, Markov AA, *et al.* Circulation in the surface and intermediate layers of the Black Sea. *Deep-Sea Research Part I*. 1993;40(8):1597-1612.
8. Mallat S. A wavelet tour of signal processing. 3rd ed. Burlington (MA): Academic Press; 2009. p. 1-805.
9. Vidakovic B. Statistical modeling by wavelets. New York: John Wiley and Sons; 1999. p. 1-382.
10. Murray JW, Top Z, Özsoy E. Hydrographic properties and ventilation of the Black Sea. *Deep-Sea Research Part A*. 1991;38(Suppl 2):S663-S689.
11. Nason GP. Wavelet methods in statistics with R. New York: Springer; 2008. p. 1-259.
12. Konovalov SK, Murray JW. Variations in the chemistry of the Black Sea on a time scale of decades (1960-1995). *Journal of Marine Systems*. 2001;31(1-3):217-243.
13. Antoniadis A, Fan J. Regularization of wavelet approximations. *Journal of the American Statistical Association*. 2001;96(455):939-967.
14. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32.
15. Capet A, Stanev EV, Beckers JM, Murray JW, Grégoire M. Decline of the Black Sea oxygen inventory. *Biogeosciences*. 2016;13(4):1287-1297.
16. Davison AC, Smith RL. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B*. 1990;52(3):393-442.
17. Kara AB, Wallcraft AJ, Hurlburt HE. Sea surface temperature sensitivity to water turbidity from simulations of the turbid Black Sea using HYCOM. *Journal of Physical Oceanography*. 2005;35(1):33-54.