Karshi Multidisciplinary International Scientific Journal

www.kmisj.reapress.com

K. Multidiscip. Int. Sci. J. Vol. 3, No. 1 (2026) 13–33.

Paper Type: Original Article

Shape-Constrained Additive Modelling of Total Petroleum Hydrocarbon Contamination: A Simulation-Based Comparison within a Gamma–Log Framework

Mirlinda Kadriaj[1], Luan Arapi[2], Raimonda Dervishi[3,*]

[1] Department of Business Informatics, Faculty of Technology Information, Tirana Business University College; mirlindakadriaj@yahoo.com.

[2] Department of Applied Geology and Geoinformatics, Faculty of Geology and Mining, Polytechnic University of Tirana, Tirana, Albania; arapiluan@gmail.com.

[3] Department of Mathematical Engineering, Faculty of Mathematical Engineering and Physical Engineering, Polytechnic University of Tirana, Tirana, Albania; raimondadervishi@yahoo.com.

Citation:

Received: 29 August 2025 *Revised: 16 October 2025* *Accepted: 08 December 2025*	Kadriaj, M., Arapi, L., & Dervishi, R. (2026). Shape-constrained additive modelling of total petroleum hydrocarbon contamination: A simulation-based comparison within a gamma–log framework. *Karshi Multidisciplinary International Scientific Journal* , *3*(1), 13-33.

Abstract

Total Petroleum Hydrocarbon (TPH) contamination poses a long-term threat to soil quality and ecosystem health in oil-producing regions. Accurate characterization of TPH drivers is challenging because environmental responses are frequently nonlinear, while available datasets are often sparse and uncertain. In this study, we evaluated whether incorporating process-informed shape constraints can improve statistical modelling of TPH concentrations. A common Gamma–log modelling framework was implemented and used to compare a Generalized Linear Model (GLM), a Generalized Additive Model (GAM), and a Shape-Constrained Additive Model (SCAM). The evaluation was performed using a site-informed simulation scenario that reflects established environmental expectations, in which TPH increases with Total Organic Carbon (TOC) and decreases with both distance from oil wells and soil depth. A synthetic dataset comprising 600 observations was generated through constrained random sampling over empirically informed covariate ranges. Model performance was examined using residual diagnostics, calibration analysis, and stratified 10-fold cross-validation. In contrast, uncertainty was assessed through Bias-Corrected and Accelerated (BCa) bootstrap intervals (B = 5,000) and local one-at-a-time sensitivity analysis (±10%). Results showed that SCAM achieved the most coherent monotonic response patterns and the strongest calibration performance, highlighting the value of integrating scientifically justified shape constraints into environmental regression models. Although conclusions remain conditional on the adopted simulation framework, the study demonstrates how constrained nonlinear models can improve interpretability and robustness in data-limited contamination assessments.

Keywords: Gamma regression, Generalized additive models, Shape-constrained additive models, Monotonicity constraints, Simulation study, Bootstrap uncertainty analysis.

1 | Introduction

The reliable statistical modelling of environmental concentration data remains challenging when empirical observations are limited, heterogeneous, and affected by substantial uncertainty. In many environmental

Corresponding Author: raimondadervishi@yahoo.com

https://doi.org/10.22105/kmisj.v3i1.115

applications, researchers must balance model flexibility against interpretability, particularly when scientific knowledge suggests directional relationships between predictors and response variables. This challenge is especially relevant for positive and right-skewed environmental variables, where the choice of functional form can substantially influence both inference and predictive performance [1–3].

Among the available modelling approaches, Gamma regression with a log link provides a natural framework for strictly positive environmental responses. Within this family, Generalized Linear Models (GLMs) offer a parsimonious and interpretable benchmark when predictor effects are approximately linear on the link scale. Generalized Additive Models (GAMs) extend this framework by allowing nonlinear relationships through smooth functions, thereby increasing flexibility while retaining a probabilistic regression structure [4], [5]. SCAMs further extend GAM methodology by incorporating scientifically motivated constraints such as monotonicity, convexity, or concavity into smooth functions [6–8]. Such constraints can improve interpretability and reduce unrealistic model behaviour when prior knowledge suggests a specific directional response.

Although GAMs and SCAMs have received increasing attention in environmental applications, comparatively few studies have examined their relative performance within controlled simulation settings where the underlying data-generating structure is known. Consequently, it remains unclear under which circumstances shape constraints provide meaningful advantages over unconstrained smooths or simpler parametric alternatives. Simulation studies provide an attractive framework for addressing this question because they allow model performance to be evaluated against a known reference structure while avoiding confounding effects introduced by measurement error, missing observations, or unknown ecological processes [9], [10].

Environmental contamination by petroleum hydrocarbons offers a suitable context for exploring these methodological questions. Total Petroleum Hydrocarbons (TPH) represent a diverse group of organic compounds commonly associated with oil extraction, processing, and transportation activities. Numerous studies indicate that TPH concentrations often exhibit structured spatial and vertical gradients, with concentrations tending to decrease as distance from contamination sources and soil depth increase, while higher organic carbon content may enhance hydrocarbon retention through sorption processes [11–17]. Although the precise functional form of these relationships remains uncertain in many field settings, their expected direction is often supported by an understanding of environmental processes.

The Patos–Marinza oilfield in Albania provides a representative environmental motivation for this study. As one of the largest onshore oilfields in Europe, Patos–Marinza has experienced decades of hydrocarbon extraction and has been associated with documented environmental pressures affecting soil and water quality [18–20]. However, publicly available datasets simultaneously describing TPH concentrations, distance-to-source relationships, and depth-dependent contamination patterns remain limited. Rather than attempting to infer site-specific environmental processes from incomplete observations, the present study uses Patos–Marinza only as inspiration for defining plausible simulation assumptions regarding the direction and relative behaviour of key predictors.

Three predictors were selected to define the baseline simulation scenario: Distance to oil wells, Total Organic Carbon (TOC), and soil depth. These variables were chosen because they consistently appear in conceptual and empirical studies of petroleum-contaminated soils and exhibit well-established directional relationships with hydrocarbon concentrations [13–17]. Within the simulation framework, TPH is assumed to increase with TOC and decrease with both distance from wells and soil depth. These relationships constitute the predefined Data-Generating Mechanism (DGM) and are not treated as empirical findings.

A central objective of environmental modelling is not only to obtain accurate predictions but also to quantify uncertainty and evaluate model robustness. Bootstrap-based inference provides a flexible framework for uncertainty assessment without strong distributional assumptions, while sensitivity analysis helps determine how conclusions respond to perturbations in model inputs and assumptions [10], [21–23]. Likewise, calibration analysis has emerged as an important complement to traditional goodness-of-fit metrics, as it

assesses whether predicted responses remain consistent with observed outcomes across the response range [24].

Against this background, this study develops a simulation-based comparative framework within a common Gamma–log regression family. Three candidate model classes are evaluated: A GLM, an unconstrained GAM, and a SCAM. By holding the distributional assumption constant and varying only the functional representation of predictor effects, the study isolates the contribution of nonlinearity and shape constraints to model performance. A synthetic dataset comprising n = 600 observations was generated through constrained random sampling over empirically informed predictor ranges, with monotonic relationships imposed through the underlying simulation mechanism.

Model performance is evaluated using residual diagnostics, calibration analysis, and stratified cross-validation. Uncertainty is quantified through Bias-Corrected and Accelerated (BCa) bootstrap confidence intervals based on 5,000 resamples, while robustness is assessed through local one-at-a-time sensitivity analysis using ±10% perturbations around the baseline scenario. All modelling approaches are evaluated under an identical simulation design to ensure a consistent comparison of predictive performance, calibration, effect recovery, and uncertainty propagation.

Significance and contributions: The contribution of this study is primarily methodological rather than environmental. Specifically, we provide a controlled comparison of GLM, GAM, and SCAM within a common Gamma–log framework and evaluate the extent to which shape constraints improve model interpretability, calibration, and robustness when the underlying signal is approximately monotonic. By combining simulation-based experimentation, bootstrap uncertainty assessment, calibration analysis, and sensitivity testing, the study establishes a transparent framework for evaluating structured nonlinear regression models under limited-data conditions. Although motivated by petroleum-contaminated soils, the proposed framework is intended to be transferable to a broader class of environmental and ecological modelling problems characterised by positive responses and scientifically motivated shape restrictions.

Objectives and hypotheses: The primary objective of this study is to compare the performance of GLM, GAM, and SCAM within a common Gamma–log framework under a controlled simulation environment. The simulation design assumes positive effects of TOC and negative effects of distance to wells and soil depth, thereby providing a known monotonic reference structure. Rather than testing these environmental relationships themselves, the study evaluates how effectively alternative model classes recover the imposed DGM.

2 | Research and Methods

2.1 | Study Area and Context

This study employs a synthetic dataset generated under a site-informed simulation framework designed to represent plausible environmental conditions encountered in petroleum-impacted regions. The simulation scenario was motivated by characteristics of the Patos–Marinza oilfield in southwestern Albania, one of the largest onshore oilfields in Europe and a region where long-term hydrocarbon extraction has been associated with documented environmental pressures [20], [25], [26].

Patos–Marinza provides an appropriate environmental context because decades of oil production, ageing infrastructure, produced-water management challenges, and enhanced oil recovery operations have created conditions under which petroleum-related contamination may occur [20], [27]. Previous environmental assessments have reported soil and water quality concerns associated with leaking wells, sludge pits, surface runoff, and waste disposal practices [25], [26]. In addition, satellite-based monitoring has identified localized ground deformation linked to long-term extraction activities [20]. These observations provide a realistic environmental motivation for the simulation design adopted in this study.

To increase ecological realism, the simulation incorporates two categorical descriptors commonly encountered in environmental monitoring programs:

Land use and soil texture: Land use is represented by four categories (agricultural, industrial, residential, and unused), reflecting broad classes of human activity that may influence contaminant exposure pathways and environmental receptors [26], [28]. Soil Texture is represented using standard USDA-style texture classes [29], acknowledging that physical soil properties may influence contaminant retention, transport, and degradation processes.

These variables were included primarily to create a more realistic modelling environment and to evaluate model behaviour under heterogeneous predictor structures. They are therefore treated as contextual covariates within the simulation rather than as empirically measured site characteristics. Because categorical variables may influence predictor distributions and model performance, cross-validation was stratified by both land use and Soil Texture to preserve class representation across training and validation subsets. This approach reduces the risk of fold-specific imbalances and provides a more stable basis for model comparison.

2.2 | Variables and Simulation Domain

A synthetic dataset was generated to provide a controlled framework for comparing alternative Gamma–log regression models under environmentally plausible conditions. The simulation scenario was informed by published studies of petroleum-contaminated soils and environmental information reported for the Patos–Marinza oilfield in Albania [11], [12], [20], [28]. The objective was not to reproduce site-specific contamination patterns, but to define realistic predictor ranges and response magnitudes within which model behaviour could be evaluated.

The response variable was (TPHs; mg/kg), simulated as a strictly positive and right-skewed variable with concentrations ranging approximately from 20 to 10,000 mg/kg. Explanatory variables included distance to oil wells (0–3,000 m), (TOC; 0–5%), soil moisture (5–35%), sampling depth (5–120 cm), soil texture (clay, loam, sand, silt), and land use (agricultural, industrial, residential, unused). Predictor ranges were informed by environmental monitoring studies and technical guidance documents relevant to petroleum-contaminated soils [13], [16], [17], [30].

The DGM assumed positive effects of TOC and negative effects of both distance to wells and soil depth, reflecting established process-based expectations reported in the literature [11], [14], [15], [28].

2.3 | Baseline Simulation Scenario and Data-Generating Mechanism

A single baseline simulation scenario was adopted to provide a transparent, reproducible, and environmentally plausible benchmark for comparing alternative Gamma–log regression models. All analyses were conducted in R 4.3.1 [31] using a fixed random seed (SEED = 123) and relevant modelling packages [5], [7]. A sample size of n = 600 was selected to ensure adequate coverage of continuous predictor domains and sufficient representation of Soil Texture and land use classes for stratified cross-validation, consistent with recommendations for simulation studies [32], [33].

Continuous predictors were generated from bounded probability distributions chosen to reflect realistic environmental variability. Distance to the nearest oil well was simulated from a truncated log-normal distribution over the interval 0–3,000 m, with approximate percentiles P10 = 165 m, P50 = 985 m, and P90 = 1,778 m [34], [35]. Sampling depth was generated from a scaled Beta distribution over 5–120 cm, yielding approximate percentiles of 10, 30, and 100 cm. TOC was generated from a scaled Beta distribution on 0–5%, with corresponding percentiles of 0.82%, 1.97%, and 3.12% [30]. Soil moisture was simulated from a scaled Beta distribution on 5–35%, with a mean of approximately 30% and a standard deviation of approximately 6.5% [11]. To represent analytical measurement uncertainty, TOC and moisture values were subjected to a small multiplicative measurement error corresponding to a Coefficient of Variation (CV) of approximately 7%.

Two categorical predictors were included to increase ecological realism. Soil Texture comprised four classes (clay, loam, sand, and silt) sampled with probabilities of 0.30, 0.30, 0.20, and 0.20, respectively [29]. Land use was represented by four categories (agricultural, industrial, residential, and unused) sampled with probabilities of 0.45, 0.20, 0.25, and 0.10 [26], [28]. These proportions were selected to represent plausible environmental conditions.

The baseline DGM imposed a monotone increase of TPH with TOC, and a monotone decrease with both distance to wells and sampling depth, while moisture was assigned only a weak effect. These directional relationships were incorporated as assumptions of the simulation design based on established environmental process understanding [11], [12], [14–17], [28]. Consequently, they should not be interpreted as empirical findings.

Response values were generated under a Gamma distribution with a log link according to

$$Y_i \sim \text{Gamma}(\mu_i, \varphi) \text{ with } \log(\mu_i) = \eta_i,$$

where μ_i denotes the conditional mean response, η_i is the linear predictor, and φ is the dispersion parameter.

The conditional mean was therefore defined as

$$\mu_i = \exp(\eta_i).$$

A dispersion value of approximately $\varphi = 0.13$ was adopted, producing realistic variability while preserving the imposed predictor–response structure. In implementation, the Gamma response was generated using shape

$$= 1/\varphi \text{ and scale} = \varphi \times \mu_i, \text{ which ensures: } E(TPH_i \mid x_i) = \mu_i \text{ and } \text{Var}(TPH_i \mid x_i) = \varphi\mu_i^2.$$

Simulated TPH concentrations were constrained to the interval 50–10,000 mg/kg to avoid unrealistic extremes while maintaining consistency with reported contamination ranges in petroleum-impacted environments [11], [12], [28]. The selected predictor domains, percentile anchors, and response bounds define a transparent and reproducible simulation framework informed by environmental evidence from petroleum-contaminated settings.

2.4 | Modelling Framework and Estimation

TPH concentrations were modelled using Gamma regression with a log link, an appropriate choice for strictly positive and right-skewed environmental response variables [1], [36]. For all models:

$$\log(\mu_i) = \eta_i,$$

where μ_i denotes the conditional mean response and η_i the linear predictor.

To isolate the contribution of functional form, three model specifications were evaluated within the same Gamma–log framework.

GLM: The GLM served as the parametric benchmark and included linear effects for distance, TOC, depth, and moisture, together with treatment contrasts for Soil Texture and land use [36].

GAM: The GAM replaced linear effects with penalised smooth functions, allowing flexible nonlinear relationships between TPH and the continuous predictors while retaining the same distributional assumptions [4], [5].

SCAM. The SCAM extended the GAM by incorporating scientifically motivated monotonicity constraints. distance and depth were constrained to have decreasing effects, TOC was constrained to have an increasing effect, and moisture remained unconstrained [6], [7].

A complete mathematical specification of the three models is provided in Appendix. The objective of the comparison was to evaluate how linear, unconstrained smooth, and shape-constrained smooth representations recover the known structure imposed by the simulation design.

2.4.1 | Estimation and diagnostic evaluation

Model estimation and diagnostics were assessed uniformly across all model specifications and follow [5] unless otherwise stated.

Residual structure: Model adequacy was evaluated using deviance and standardised residuals plotted against fitted values, Quantile–Quantile (QQ) plots of deviance residuals, and scale–location diagnostics to assess residual spread across the fitted range.

Smooth sufficiency: Basis adequacy was assessed using the k-index diagnostic and associated p-values to identify potentially under-dimensioned spline bases. These diagnostics were evaluated primarily for the unconstrained GAM and, where compatible, for the SCAM using identical smooth-term specifications.

Complexity and identifiability: Smooth complexity was quantified through Effective Degrees of Freedom (EDF). For the GAM, nonlinear dependence among smooth terms was assessed using the concurvity measures implemented in mgcv. Because no native concurvity diagnostic is available for SCAM, an exploratory concurvity proxy was computed as the coefficient of determination (R^2) obtained by regressing each smooth term's contribution to the linear predictor against the contributions of the remaining smooth terms. Higher values indicate greater nonlinear collinearity among smooth components [6].

Calibration assessment: Calibration was evaluated by regressing observed log (TPH) values on the predicted linear predictor ($\hat{\eta}$), with ideal calibration corresponding to an intercept of 0 and a slope of 1. Estimates, standard errors, confidence intervals, and p-values were reported for both coefficients. In addition, smooth calibration curves with loess overlays were inspected to identify potential local departures from calibration across the prediction range [24].

Dispersion and residual behaviour: Model dispersion was evaluated using Gamma dispersion estimates obtained from model summaries, together with Pearson residual summaries and exceedance rates $Pr(|r_p| > c)$, where r_p denotes the Pearson residual and c represents a predefined diagnostic threshold [1], [36].

Because the baseline DGM incorporates monotone relationships that are broadly compatible with the nonlinear regressions, model performance is interpreted in terms of calibration, diagnostic behaviour, and recovery of the imposed structure within the assumed simulation scenario.

2.5 | Cross–Validation and Model Comparison

We used stratified 10-fold cross-validation, stratifying on the Soil Type × land use combination to preserve subgroup representation across folds. The same fold assignments were used for the GLM**,** GAM, and SCAM to ensure a fair comparison. Out-of-sample performance was summarized using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), mean per-observation deviance, and mean log-likelihood. In-sample model summaries included AIC (ML), deviance explained (%), and the log-likelihood. By holding the distributional family and link constant (Gamma–log) across models, observed performance differences can be attributed primarily to differences in functional form (linear vs. smooth vs. constrained smooth) rather than to the assumed error structure [5], [37], [38].

2.6 | Sensitivity and Uncertainty Analysis

To evaluate the robustness of model predictions and the stability of estimated effects, local sensitivity analysis and bootstrap-based uncertainty quantification were conducted for the SCAM model. This additional analysis was performed after the comparative evaluation of GLM, GAM, and SCAM because SCAM provided the most coherent representation of the monotonic structure imposed in the baseline DGM. Consequently,

uncertainty assessment focused on the model selected for detailed interpretation within the simulation framework.

Local sensitivity analysis: Local sensitivity analysis was performed using a One-at-a-Time (OAT) perturbation design [9], [10]. Continuous predictors selected for interpretation in the fitted SCAM (distance to wells, TOC, and depth) were individually perturbed by ±10% around a common baseline defined by median predictor values. During each perturbation, all remaining continuous predictors were held at their median values and categorical predictors at their reference levels. For each perturbation, changes in the expected response (μ) and corresponding local elasticities were computed relative to the baseline condition. This approach provides an interpretable assessment of predictor influence while avoiding confounding effects associated with simultaneous perturbation of multiple inputs [9], [10].

Bootstrap resampling: We quantified uncertainty via nonparametric bootstrap (B = 5,000) by resampling rows (sites) with replacement and refitting the SCAM on each resample (same monotonicity constraints; REML smoothing). For every refit, we recomputed: 1) scenario means for eight decision–relevant combinations crossing TOC (low p10 / high p90), distance (near p10 / far p90), and depth (shallow P10 / deep P90), and 2) centered partial effects at the 10th/50th/90th percentiles (others predictors held at baseline). We formed 95% BCa intervals using jackknife–based bias correction and acceleration, with a percentile fallback if BCa was unavailable [21–23]. Resampling at the row level preserves the joint covariate structure for scenario predictions. We report the bootstrap mean and BCa interval alongside the non–resampled model mean to assess bootstrap bias. Uncertainty bands are shown on the response scale (mg/kg) for scenarios and on the link scale for partial effects.

3 | Results

3.1 | Model Coefficients and Performance Comparison

Covariate effects and functional forms: The Gamma–log regression models – GLM, GAM, and SCAM – yielded broadly consistent estimates across specifications, with convergence in both the direction and magnitude of effects. *Table 1* summarises areas of agreement as well as model specifications differences in predictor effects. Intercept terms were statistically significant in all three models. However, these results should be interpreted primarily as model constants under the chosen parameterisation rather than as substantively meaningful estimates of contamination independent of the covariates.

Organic carbon (TOC) was the most robust predictor. In the GLM, TOC was strongly positive and statistically significant ($\beta = 0.160$, $p < 0.001$), implying that each one unit increase in TOC is associated with ≈17% higher expected TPH ($\exp(0.160) \approx 1.174$). The GAM and SCAM smooths supported the same positive direction: TOC only limited nonlinearity in the GAM (edf = 1.949, $p < 0.001$) and was essentially linear in the SCAM (edf ≈ 1.000, $p < 0.001$), consistent with the baseline simulation design.

Spatial and vertical covariates also exhibited clear patterns. Distance to oil well had a significant negative effect in the GLM ($p < 0.001$), while the GAM (edf = 2.926, $p < 0.001$) and SCAM (edf = 2.183, $p < 0.001$) confirmed nonlinear decreases in TPH with increasing distance. Similarly, depth was consistently negative (GLM: $\beta = -0.006$, $p < 0.001$), implying ≈0.6% lower expected TPH per centimetre; GAM/SCAM smooths again supported gentle monotone declines (edf = 2.530 and 2.101; both $p < 0.001$). These results are directionally coherent with the adopted monotone baseline scenario, in which TPH was generated to decline with both source distance and sampling depth. In contrast, moisture was not statistically significant at the 0.001 level in any specification (GLM: $p = 0.070$; GAM: $p = 0.088$; SCAM: $p = 0.077$), suggesting limited independent contribution after accounting for TOC, soil type, and spatial gradients. While moisture can influence hydrocarbon degradation, its effect is often contingent on microbial and site conditions.

For soil texture (reference = clay), sand was consistently negative and statistically significant (GLM: $\beta = -0.117$, $p = 0.003$; GAM: $\beta = -0.121$, $p = 0.002$; SCAM: $\beta = -0.123$, $p = 0.002$), whereas loam and silt were not significant. Regarding land use (reference = agricultural), industrial areas showed modest positive

associations (GLM: β = 0.084, p = 0.034; GAM: p = 0.056; SCAM: p = 0.053), residential areas were not significant (GAM: p = 0.469; SCAM: p = 0.416), and unused land showed negative associations (GLM: β = −0.135, p = 0.010; GAM: β = −0.149, p = 0.004; SCAM: β = −0.143, p = 0.006). Overall, TOC, soil texture, distance, and depth appear more influential than land-use categories within the baseline simulated scenario.

Table 1. Estimated coefficients and significance levels for Gamma, log link regression models under the baseline simulated scenario. Reference levels are clay soil for soil type and agricultural land for land use. EDF are reported for GAM and SCAM smooth terms.

Variables	GLM	GAM	SCAM
Parametric term			
Intercept	+5.961 (p<0.001)	+5.644 (p<0.001)	+5.644 (p<0.001)
Soil texture: Loam	-0.029 (p=0.437)	-0.034 (p=0.351)	−0.035 (p=0.346)
Soil texture: Sand	-0.117 (p=0.003)	-0.121 (p=0.002)	−0.123 (p=0.002)
Soil texture: Silt	-0.067 (p=0.131)	-0.059 (p=0.174)	−0.062 (p=0.153)
Land use: Industrial	+0.084 (p=0.034)	+0.074 (p=0.056)	+0.074 (p=0.053)
Land use: Residential	+0.040 (p=0.270)	+0.025 (p=0.469)	+0.029 (p=0.416)
Land use: Unused	-0.135 (p=0.010)	-0.149 (p=0.004)	-0.143 (p=0.006)
Continues predictors			
Distance to well (m)	-0.0002 (p<0.001)	edf= 2.926 (p<0.001)	edf=2.183 (p<0.001)
Organic carbon (%)	+0.160 (p<0.001)	edf= 1.549 (p<0.001)	edf=0.998 (p<0.001)
Moisture (%)	-0.004 (p=0.070)	edf= 0.088 (p=0.699)	edf=0.000 (p=0.077)
Depth (cm)	-0.006 (p<0.001)	edf= 2.530 (p<0.001)	edf=2.101 (p<0.001)

Note: Reference level for categorical predictors is agricultural land and clay soil.

Model comparison (Gamma log) and cross–validation. As summarized in *Table 2*, the SCAM achieved the lowest AIC (7106), outperforming the GAM (7111; ΔAIC = 4.250) and the GLM (7137; ΔAIC = 30.200). The GAM obtained the highest log-likelihood (−3538) and a slightly higher fraction of deviance explained (47.6%) than SCAM (−3539; 47.4%), but after accounting for model complexity the AIC criterion favored SCAM. Out-of-sample performance from cross-validation showed modest yet consistent advantages for SCAM: CV-RMSE was tied for GAM and SCAM (105) and better than GLM (107); CV-MAE was lowest for SCAM (76.9 vs 77.8 for GAM and 79.1 for GLM); CV-Deviance was lowest for SCAM (0.122 vs 0.123 and 0.128); and CV-LogLik was least negative for SCAM (−5.920 vs −5.930 and −5.950). These differences are modest in magnitude, but they are directionally consistent across several criteria. Under the baseline monotone simulated scenario considered in the study, they suggest that SCAM provided the best balance between fit, calibration, and structural parsimony.

Table 2. Model comparison under Gamma–log models (GLM, GAM, SCAM) for the baseline simulated scenario.

Model	AIC (ML)	ΔAIC	Log Lik (ML)	Deviance Explained (%)	CV-RMSE	CV-MAE	CV-Deviance	CV-LogLik
GLM	7137.000	30.200	-3556.000	44.300	107.000	79.100	0.128	-5.950
GAM	7111.000	4.250	-3538.000	47.600	105.000	77.800	0.123	-5.930
SCAM	7106.000	0.000	-3539.000	47.400	105.000	76.900	0.122	-5.920

Notes: Cross-validation is 10-fold and stratified by soil texture × land use. "Better" values are lower for AIC, CV-RMSE, CV-MAE, and CV-deviance; higher for Log Lik and CV-LogLik (mean per-observation Log-Likelihood). Deviance explained (%) is computed on the full dataset. For GAM, AIC/Log Lik are taken from an ML refit; for SCAM, the reported AIC/log-likelihoods are from the penalised fit.

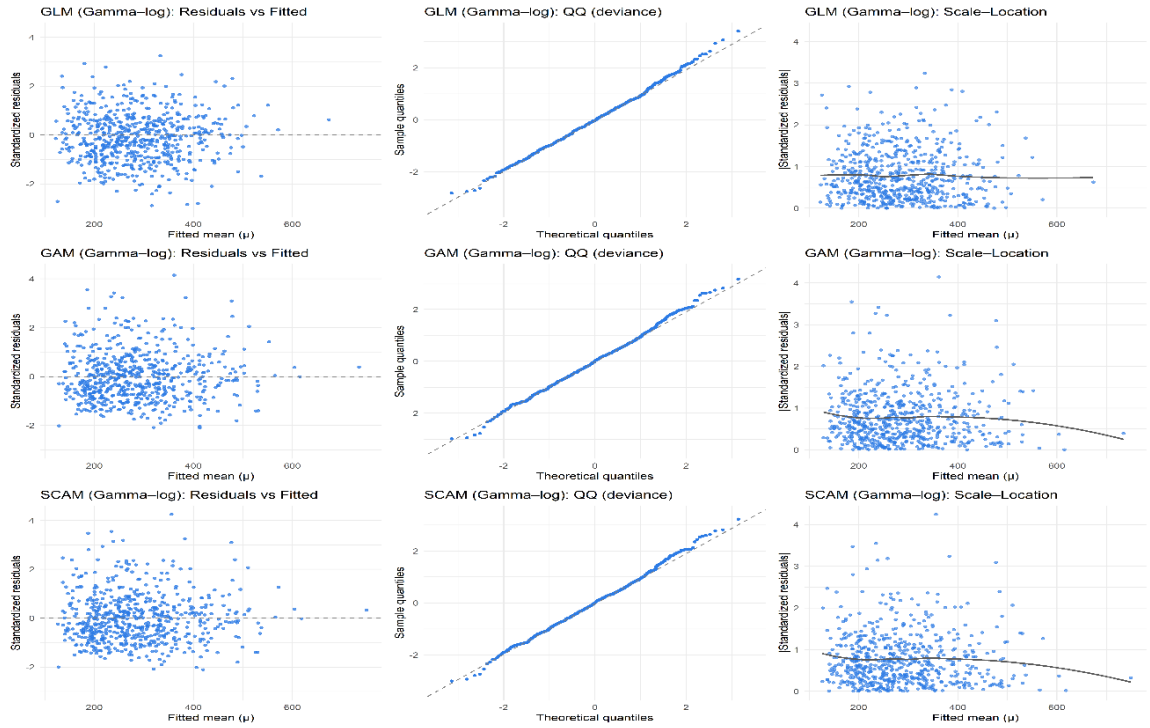

Fig. 1. Diagnostic panels for Gamma–log models. Rows correspond to the three fitted models—GLM (Gamma–log), GAM (Gamma–log), and SCAM (Gamma–log)—and columns show; a. residuals vs fitted (standardized residuals with zero reference line), b. QQ-plot of deviance residuals (with reference line), and c. scale–location plot (standardized residuals r* vs fitted μ).

Residual diagnostics, dispersion, and calibration. Diagnostic panels (*Fig. 1*) showed no systematic structure in residuals across models; QQ-plots of deviance residuals adhered closely to the reference line, and scale–location plots indicated stable dispersion over the fitted mean with only minor widening at the extremes, a pattern consistent with a Gamma–log mean–variance relation. Correspondingly, dispersion estimates decreased from φ=0.125 (GLM) to 0.119 (GAM, SCAM). Distributions of residuals were well-behaved (e.g., q95(|rPearson|)=0.683, 0.693, 0.692; q99(|rPearson|)= 1.076, 1.087, 1.070 for GLM, GAM, SCAM, respectively), and the proportion of observations with absolute Pearson residuals greater than 2 or greater than 3 was 0.00% for all models. Calibration was strong across specifications: calibration intercepts were close to zero (−0.061, −0.118, −0.038) and slopes close to one (1.000, 1.011, 0.996) with non–significant tests of H_0: Slope = 1 (*Table 3*).

Table 3. Residual diagnostics, dispersion, and calibration under Gamma–log models (GLM, GAM, SCAM) for the baseline simulated scenario.

Metric	GLM	GAM	SCAM
Dispersion ϕ	0.125	0.119	0.119
Deviance resid. mean	−0.041	−0.038	−0.038
Deviance resid. sd	0.348	0.338	0.339
Pearson resid. sd	0.353	0.341	0.342
Pr($\lvert r_{Pearson} \rvert > 2$) (%, over n=600)	0.00% (0)	0.00% (0)	0.00% (0)
Pr($\lvert r_{Pearson} \rvert > 3$) (%, over n=600)	0.00% (0)	0.00% (0)	0.00% (0)
q95 ($\lvert r_{Pearson} \rvert$)	0.683	0.693	0.692
q99 ($\lvert r_{Pearson} \rvert$)	1.076	1.087	1.070
Calibration: Intercept	−0.061	−0.118	−0.038
Calibration: Slope	1.000	1.011	0.996
Smooth diagnostics			
edf (distance)	NA	2.929	2.183
edf (depth)	NA	2.530	2.101
k-index (distance, p)	NA	0.983 (p=0.453)	0.982 (p=0.448)

Table 3. Continued.

Metric	GLM	GAM	SCAM
k-index (depth, p)	NA	0.914 (p=0.035)	0.911 (p=0.022)
Concurvity (distance)	NA	1.0000	0.0095 (proxy)
Concurvity (depth)	NA	1.0000	0.0077 (proxy)

Notes: φ = Gamma dispersion under log link; $r_{Pearson}$ = Pearson residual; Pr ($|r_{Pearson}| > c$) is computed over n = 600; edf = EDF; k-index = spline basis sufficiency diagnostic; concurvity for SCAM is reported as a proxy R^2 obtained by regressing each smooth's linear predictor component on the remaining smooth components.

Nonlinear effects and smooth diagnostics. Distance and depth exhibited gentle nonlinearity (GAM edf = 2.93, 2.53; SCAM edf = 2.18, 2.10), whereas TOC was near–linear (GAM edf = 1.949; SCAM edf ≈ 1.000). Basis–sufficiency was adequate for distance (GAM: k–index = 0.983, p = 0.453; SCAM: k–index = 0.982, p = 0.448) and suggested mild under–dimensioning for depth (GAM: k–index = 0.914, p = 0.035; SCAM: k–index = 0.911, p = 0.022), though the fitted trends remained stable and no material change in substantive conclusions was observed. Concurvity proxies for SCAM were near zero (distance = 0.0095; depth = 0.0077); the GAM observed concurvity metric equals 1.000 by construction and is not directly comparable (*Table 3*). In SCAM, three covariates remained FDR significant—distance to oil wells, TOC, and depth—with monotone, interpretable partial effects on log(TPH) and pointwise 95% intervals (*Figu. 2*): expected TPH decreased with distance and depth and increased with TOC.

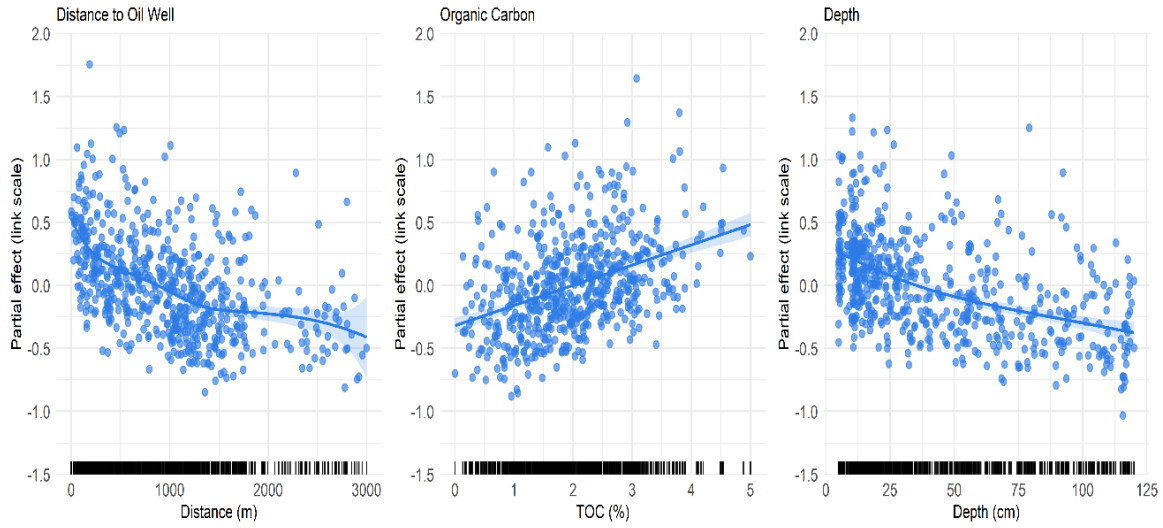

Fig. 2. SCAM (Gamma, log link) partial effects for distance to oil well and depth (shape-constrained smooths); TOC enters linearly. Points show partial residuals; lines show estimated effects with pointwise 95% intervals (±2 SE) on the log(TPH) scale.

3.2 | Sensitivity and Uncertainty Analysis

Sensitivity and bootstrap-based uncertainty analyses were conducted for SCAM as the primary scenario-interpretation model, after all three candidate models had first been compared using the same Gamma–log framework, diagnostics, calibration, and cross-validation procedures. These additional analyses are therefore intended to characterize uncertainty within the constrained baseline simulation scenario, rather than to serve as the sole basis for ranking model classes. Sensitivity analysis was restricted to the covariates that were FDR–significant in the SCAM (distance to wells, TOC, depth); all other covariates were held at their baseline values. At the common baseline (distance = 983 m, depth = 30 cm, TOC = 1.97%), one-at–a–time ±10% perturbations indicate directionally coherent and locally modest effects on predicted TPH under the SCAM

(Gamma–log) model (*Table 4*). Shortening the distance to wells by 10% increases mean TPH by 4.1% (absolute +11.2 mg/kg), whereas lengthening the distance by 10% decreases it by 3.6% (−9.8 mg/kg), implying a central elasticity of approximately −0.382 (i.e., a 1% increase in distance reduces the expected mean μ by ~0.38%). Soil depth shows a weaker negative influence: Making soils 10% shallower raises TPH by 2.4% (+6.6 mg/kg), while making soils 10% deeper lowers it by 2.2% (−6.1 mg/kg), corresponding to an elasticity of approximately −0.218. In contrast, TOC exerts a positive effect: A 10% reduction in TOC lowers TPH by 3.1% (−8.5 mg/kg) and a 10% increase in TOC raises it by 3.2% (+8.8 mg/kg), with elasticity approximately +0.327. These perturbation results are internally coherent with the baseline simulation design and indicate that, around the median baseline condition, distance and TOC have the largest local influence on predicted TPH, while depth has a smaller but still directionally consistent effect.

Table 4. Local sensitivity analysis (±10%) for the SCAM (Gamma–log) under the baseline simulated scenario. Percentage change in predicted soil TPH (response scale, mg/kg) is reported relative to the common baseline while all non-target covariates are held fixed.

Covariate	Distance to Oil Well	Soil Depth	TOC
Baseline value	983.0	30.0	1.97
−10% value	884.0	27.0	1.77
+10% value	1081.0	33.0	2.17
Baseline μ	275.0	275.0	275.0
μ (−10%)	286.0	281.0	266.0
Absolute change (−10%) (mg/kg)	+11.2	+6.6	−8.5
Relative change (−10%)	+4.1	+2.4	−3.1
μ (+10%)	265.0	269.0	284.0
Absolute change (+10%) (mg/kg)	−9.8	−6.1	+8.8
Relative change (+10%)	−3.6	−2.2	+3.2
Central elasticity (baseline)	−0.382	−0.218	+0.327

Note: Baseline uses median values for continuous covariates and reference levels for factors. This analysis is local and descriptive; it does not represent an intervention analysis for the real field system.

Fig. 3 displays the local ±10% perturbations as dumbbell plots around the common baseline under the SCAM (Gamma–log) model. Responses are directionally coherent and only mildly nonlinear: shortening the distance to wells increases mean TPH, increasing depth lowers it, and increasing TOC raises it. Magnitudes are small to moderate at the baseline (single–digit percent changes), with distance and TOC exerting the largest local influence and depth the smallest, an ordering consistent with the SCAM partial–effect results and with physical expectations. The slight asymmetry between −10% and +10% shifts is consistent with curvature in the fitted constrained smooths rather than with instability in the model.

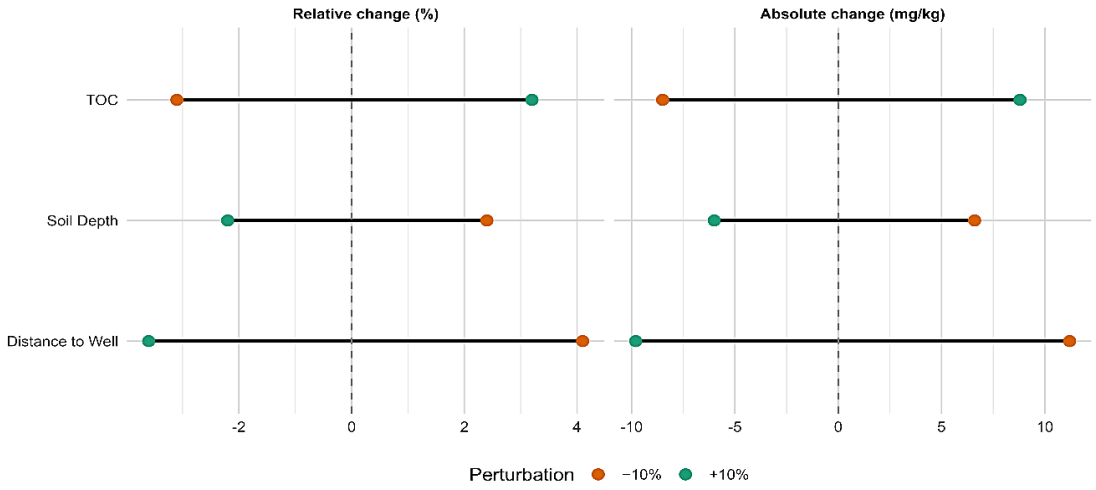

Fig. 3. Local sensitivity analysis (±10%) for the SCAM (Gamma–log) around a common baseline (left: relative change (%); right: absolute change (mg/kg)).

To quantify uncertainty in the fitted SCAM, we used nonparametric bootstrap resampling (B=5,000) with BCa 95% confidence intervals. Smooth effects were evaluated on the link scale ($\eta = \log \mu$) at the 10th, 50th, and 90th percentiles of each covariate, holding other predictors at baseline. Positive deviations indicate higher log(TPH) relative to the median condition. The common baseline at P50 is η approximately 5.62 (95% BCa: 5.5292–5.7241), as shown in *Fig. 4* and *Table 5*.

Table 5. SCAM (Gamma–log) partial effects at covariate quantiles. Estimated effects are reported on the link scale ($\eta = \log \mu$), evaluated at the 10th/50th/90th percentiles, with other continuous predictors at their medians and factors at reference levels. Intervals are BCa 95% bootstrap confidence intervals (B=5,000).

Covariate	EDF	Level	Covariate value	Estimate (Log)	95% CI (BCa)
Distance to oil well	4.5	P10	163.00	5.99	[5.9013, 6.1177]
Distance to oil well	4.5	P50	983.00	5.62	[5.5292, 5.7241]
Distance to oil well	4.5	P90	1777.00	5.46	[5.3739, 5.5613]
Soil depth	2.2	P10	9.94	5.79	[5.6998, 5.8762]
Soil depth	2.2	P50	30.00	5.62	[5.5292, 5.7241]
Soil depth	2.2	P90	100.00	5.28	[5.1772, 5.3859]
TOC	1.0	P10	0.82	5.43	[5.3412, 5.5448]
TOC	1.0	P50	1.97	5.62	[5.5292, 5.7241]
TOC	1.0	P90	3.12	5.80	[5.7020, 5.9167]

Note: P50 rows coincide with the common baseline. Approximate percent changes on the response scale can be obtained as $\exp(\Delta\log)-1$. The reported sign-stability values indicate that the fitted monotone directions were preserved across bootstrap resamples within the baseline scenario.

Distance to oil wells shows the expected near-field elevation in TPH. At the 10th percentile of distance (approximately 163 m), η is approximately 5.99 (95% BCa: 5.9013–6.1177), which is about 0.37 higher than the median and corresponds to an estimated increase of roughly 45% on the response scale. At the 90th percentile of distance (approximately 1,777 m), η is approximately 5.46 (95% BCa: 5.3739–5.5613), or about 0.16 lower than the median, equivalent to an estimated decrease of about 15%. Soil depth exhibits a similar attenuation pattern. At shallow sampling depths (around 10 cm), η is approximately 5.79 (95% BCa: 5.6998–5.8762), which is about 0.17 above the median, corresponding to an increase of roughly 19%. At deeper sampling depths (around 100 cm), η is approximately 5.28 (95% BCa: 5.1772–5.3859), which is about 0.34 below the median, corresponding to an estimated decrease of roughly 29%. TOC shows a strong increasing effect. At low TOC (about 0.82%), η is approximately 5.43 (95% BCa: 5.3412–5.5448), which is about 0.19 below the median (an estimated decrease of roughly 17%). At high TOC (about 3.12%), η is approximately 5.80 (95% BCa: 5.7020–5.9167), which is about 0.18 above the median (an estimated increase of roughly 20%). Across all three covariates, sign-stability was 100%, which supports the internal consistency of the constrained smooths under resampling within the adopted baseline scenario.

Bootstrap–based mean predictions for eight representative scenarios (*Table 6* and *Fig. 5*) confirm a coherent and physically plausible ordering of TPH on the response scale. Across all combinations of distance, depth, and TOC, high–TOC scenarios produce higher predicted TPH than low–TOC scenarios; locations near wells produce higher TPH than locations far from wells; and shallow soils produce higher TPH than deeper soils. Bootstrap means range from 142 mg/kg (low TOC, far from wells, deep sampling depth; 95% BCa: 126.0–158.2 mg/kg) to 568 mg/kg (high TOC, near wells, shallow depth; 95% BCa: 502.0–641.2 mg/kg).

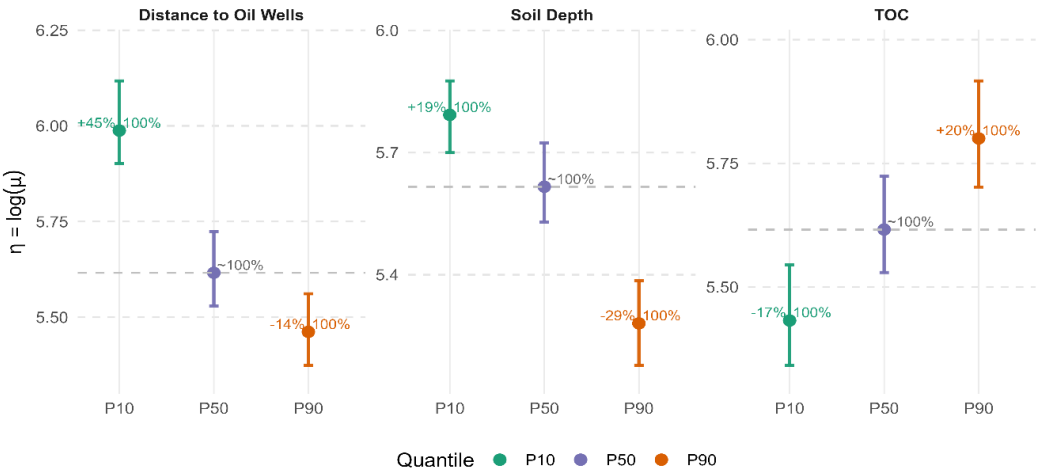

Fig. 4. SCAM (Gamma–log) partial effects at the 10th/50th/90th percentiles of distance to oil wells, soil depth, and TOC under the baseline simulated scenario. Effects are shown on the link scale (η = log μ); vertical whiskers denote BCa 95% CIs (B = 5,000; percentile fallback where BCa was unavailable).

Table 6. Bootstrap-based mean predictions of soil TPH (mg/kg) for eight representative scenarios combining TOC (low ≈ P10, high ≈ P90), distance to oil wells (near ≈ P10, far ≈ P90), and soil depth (shallow ≈ P10, deep ≈ P90).

Scenario	TOC (%)	Distance (m)	Depth (cm)	Predicted μ (mg/kg)	Bootstrap Mean μ	95% CI (μ)
Low TOC, near, shallow	0.82	163.0	10	396	390	[346.2, 435.3]
Low TOC, near, deep	0.82	163.0	100	237	236	[205.6, 268.0]
Low TOC, far, shallow	0.82	1777.0	10	234	234	[211.2, 258.8]
Low TOC, far, deep	0.82	1777.0	100	140	142	[126.0, 158.2]
High TOC, near, shallow	3.12	163.0	10	572	568	[502.0, 641.2]
High TOC, near, deep	3.12	163.0	100	342	343	[295.7, 393.5]
High TOC, far, shallow	3.12	1777.0	10	338	341	[305.6, 377.4]
High TOC, far, deep	3.12	1777.0	100	202	206	[181.8, 232.8]

Note: Other continuous covariates were fixed at their medians, and factors were set at their reference levels. Values are on the response scale (μ), intervals are BCa 95% across bootstrap refits (B = 5,000).

Fig. 5 visualizes these eight scenarios. Filled points represent bootstrap means with their BCa 95% confidence intervals, and open points represent model means obtained without resampling. In all cases, the model means fall within the corresponding bootstrap intervals. The absolute differences between the model means and the bootstrap means are small (≤ 6 mg/kg, corresponding to ≤ 2%), indicating negligible bootstrap bias and stable estimates. The proportional contrasts are substantial and internally consistent with a multiplicative Gamma–log mechanism. Increasing TOC from approximately 0.82% to approximately 3.12% increases the predicted mean TPH by roughly 45% in every distance × depth combination (e.g., from 234 to 341 mg/kg, from 142 to 206 mg/kg, from 390 to 568 mg/kg, and from 236 to 343 mg/kg). Increasing sampling depth from about 10 cm to about 100 cm lowers the predicted mean TPH by roughly 39% to 40% (e.g., from 390 to 236 mg/kg, from 568 to 343 mg/kg, from 234 to 142 mg/kg, and from 341 to 206 mg/kg). Moving from near wells (around 163 m) to far from wells (around 1,777 m) lowers the predicted mean TPH by roughly 40% under otherwise comparable conditions (e.g., from 390 to 234 mg/kg and from 343 to 206 mg/kg). These response-scale contrasts provide interpretable scenario summaries under the constrained baseline design.

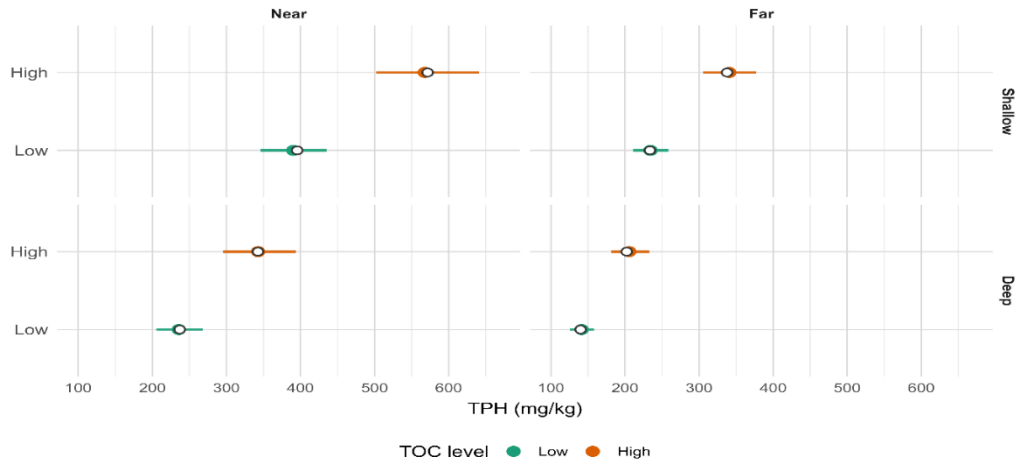

Fig. 5. Forest plot of bootstrap-based mean predictions (B = 5,000) for eight SCAM-based scenarios crossing TOC (low/high), distance to oil wells (near/far), and soil depth (shallow/deep).

On the log–link scale, these contrasts correspond to shifts of approximately the same magnitude reported in the smooths. Moving from near wells (~163 m) to far wells (~1,778 m) reduces the linear predictor by about 0.53 ($\Delta\eta \approx -0.53$), which corresponds to an approximate 41% decrease in the mean on the response scale. Increasing depth from shallow (~10 cm) to deep (~100 cm) reduces the linear predictor by about 0.51 ($\Delta\eta \approx -0.51$), or roughly a 40% decrease on the response scale. Increasing TOC from ~0.82% to ~3.12% raises the linear predictor by about 0.37 ($\Delta\eta \approx +0.37$), which corresponds to an increase of about 45% in the mean. The consistency between link-scale contrasts, response-scale contrasts, and bootstrap ordering suggests that the SCAM captured the imposed baseline gradients in a clear and interpretable manner within the adopted simulation framework.

4 | Discussion

This study demonstrates how a simulation-based framework can be used to evaluate alternative Gamma–log regression models under controlled conditions where the underlying DGM is known. Rather than attempting to infer environmental processes from observational data, the objective was to compare how effectively GLM, GAM, and SCAM recover a predefined monotonic structure motivated by environmental process understanding. Within the adopted simulation framework, TOC, distance to oil wells, and soil depth constituted the dominant components of the imposed signal, whereas soil texture, land use, and moisture contributed comparatively little explanatory information. Across all model specifications, effect directions were recovered consistently, with positive effects of TOC and negative effects of both distance and depth. SCAM provided the clearest representation of these relationships, together with the strongest calibration, the most favourable information criteria, and modest but consistent improvements in cross-validated predictive performance.

Methodologically, the comparison highlights the progressive trade-off between simplicity, flexibility, and structural knowledge. The GLM provided a transparent benchmark but was limited in its ability to represent nonlinear behaviour [1], [36]. The GAM improved model flexibility by allowing smooth nonlinear effects and produced systematic gains in fit and predictive accuracy [4], [5]. SCAM further incorporated prior structural information through monotonicity constraints, yielding effect profiles that were both scientifically coherent and highly interpretable [6], [7]. The baseline DGM contained monotonic relationships, consequently, the results should be interpreted as a demonstration of the potential benefits of shape constraints when reliable prior knowledge about effect direction is available [5], [6].

Sensitivity and bootstrap analyses provided an additional assessment of model stability. Local one-at-a-time perturbations, bootstrap partial effects, and scenario-based predictions produced highly consistent results, indicating that the fitted SCAM stably recovered the imposed monotonic structure. The agreement between

sensitivity measures, bootstrap confidence intervals, and scenario predictions provides an internal consistency check for the modelling framework and supports the robustness of the resulting interpretations. Because these analyses were performed on a synthetic dataset, however, they quantify stability within the adopted simulation framework rather than uncertainty associated with real-world contamination processes [9], [10], [21], [22].

A limitation of this study is that the comparison was conducted under a single site-informed simulation scenario in which monotonic relationships were embedded directly within the DGM. Consequently, the observed advantages of SCAM should be interpreted within the context of the assumed simulation framework rather than as evidence of universal superiority across all environmental applications. In addition, the analysis did not explicitly model spatial dependence or temporal variability, and the findings therefore reflect model behaviour under controlled conditions rather than real-world system complexity. Future work should extend the framework to alternative simulation scenarios, including non-monotonic relationships, heterogeneous uncertainty, and observational datasets, to evaluate the robustness and generality of the conclusions.

By holding the distributional assumption constant and varying only the functional representation of predictor effects, the study isolates the marginal value of nonlinearity and shape constraints within a common Gamma–log framework. The results suggest that SCAMs can improve calibration, interpretability, and recovery of known monotonic structures while maintaining competitive predictive performance. These findings are consistent with broader arguments favouring transparent and interpretable modelling approaches when scientific understanding, accountability, and decision support are important considerations [39], [40]. More broadly, the proposed workflow provides a transparent template for comparing structured regression models in settings where empirical data are limited but scientifically motivated assumptions about effect direction are available.

Similar simulation-based comparisons could be applied to a broad range of environmental, ecological, and geoscientific problems involving positive responses, nonlinear effects, and prior expectations regarding monotonic behaviour. Any operational application, however, would require validation and recalibration using observed field data before site-specific conclusions could be drawn. Consequently, the findings should be interpreted as evidence regarding model behaviour under a controlled simulation environment rather than as direct evidence about contamination dynamics in a particular field setting.

5 | Conclusion

This study developed a simulation-based framework for comparing Gamma–log GLM, GAM, and SCAM models under a common DGM. By holding the distributional assumption constant and varying only the functional representation of predictor effects, the framework enabled a direct assessment of the contribution of linear, unconstrained nonlinear, and shape-constrained nonlinear modelling strategies.

The results showed that all three approaches recovered the imposed directional structure, but SCAM provided the most coherent representation of the assumed monotonic relationships, together with favourable calibration, information criteria, and predictive performance. These findings suggest that shape constraints can be beneficial when reliable prior knowledge supports monotonic behaviour in the underlying process.

The main contribution of this work is the development of a transparent methodological framework for evaluating structured regression models under limited-data conditions. Although motivated by petroleum-contaminated soils, the proposed approach is transferable to a broader range of environmental and geoscientific applications involving positive responses, nonlinear effects, and scientifically justified directional assumptions. Future work should evaluate the framework under alternative simulation scenarios and observational datasets to assess the generality of the conclusions.

Acknowledgments

Not applicable.

Author Contributaion

Conceptualization, K. M., and A. L.; methodology, K. M., and D. R.; software, A. L.; validation, A. L., K. M., and D. R.; formal analysis, De. R.; investigation, A. L., and K. M.; writing-creating the initial design, A. L.; writing-reviewing and editing, K. M.; visualization, K. M.; monitoring, K. M., and A. L. All authors have read and agreed to the published version of the manuscript.

Data Availability

The study is based on theoretical simulation; no new data were created. The simulated code is available from the corresponding author upon reasonable request.

Funding

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Consent for Publication

The author confirms consent for the publication of this work

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

[1] Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315182780

[2] Helsel, D. R. (2011). *Statistics for censored environmental data using Minitab and R*. John Wiley & Sons. https://doi.org/10.1002/9781118162729

[3] Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R* (Vol. 574, p. 574). New York: Springer. https://doi.org/10.1007/978-0-387-87458-6

[4] Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical science*, *1*(3), 297–310. https://doi.org/10.1214/ss/1177013604

[5] Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

[6] Pya, N., & Wood, S. N. (2015). Shape constrained additive models. *Statistics and computing*, *25*(3), 543–559. https://doi.org/10.1007/s11222-013-9448-7

[7] Pya, N. (2016). *Scam: Shape constrained additive models (R package version 1.x).* https://cran.r-project.org/package=scam

[8] Arnqvist, N. P. (2024). *On some extensions of shape-constrained generalized additive modelling in R*. https://doi.org/10.48550/arXiv.2403.09438

[9] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., ... & Tarantola, S. (2008). *Global sensitivity analysis: The primer*. John Wiley & Sons. https://doi.org/10.1002/9780470725184

[10] Iooss, B., & Lemaître, P. (2015). A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems: Algorithms and applications* (pp. 101–122). Springer. https://doi.org/10.1007/978-1-4899-7547-8_5

[11] Registry., A. for T. S. and D. (1999). *Toxicological profile for total petroleum hydrocarbons (TPH)*. https://www.atsdr.cdc.gov/toxprofiles/tp123.pdf

[12] Canadian Council of Ministers of the Environment. (2008). *Canada-wide standard for petroleum hydrocarbons (PHC) in soil: Scientific rationale: Supporting technical document*. https://ccme.ca/en/res/cws_phc_user_guide_1.1_e.pdf

[13] Doucette, W. J. (2003). Quantitative structure-activity relationships for predicting soil-sediment sorption coefficients for organic chemicals. *Environmental toxicology and chemistry*, *22*(8), 1771–1788. https://doi.org/10.1897/01-362

[14] Palma, P., Fialho, S., Alvarenga, P., Santos, C., Brás, T., Palma, G., ... & Neves, L. A. (2016). Membranes technology used in water treatment: Chemical, microbiological and ecotoxicological analysis. *Science of the total environment*, *568*, 998-1009. https://doi.org/10.1016/j.scitotenv.2016.04.208

[15] Karickhoff, S. W., Brown, D. S., & Scott, T. A. (1979). Sorption of hydrophobic pollutants on natural sediments. *Water research*, *13*(3), 241–248. https://doi.org/10.1016/0043-1354(79)90201-X

[16] Yu, L., Wang, J., Li, Y., & Chen, B. (2019). Distribution and migration of petroleum hydrocarbons around oil wells in contaminated soils. *Environmental monitoring and assessment*, *191*(502), 13. https://doi.org/10.1007/s10661-019-7659-9

[17] Zhu, L., Li, W., & Pan, L. (2015). Horizontal and vertical migration of petroleum hydrocarbons in oilfield soils. *Environmental science: Processes & impacts*, *17*(3), 659–667. https://doi.org/10.1039/C4EM00628J

[18] Seiti, B., Topi, D., Lame, A., & Drushku, S. (2010). The evaluation of environmental situation due to the oil industry activity in Albania. *BALWOIS 2010 conference, Ohrid, North Macedonia* (p. 469). BALWOIS / Balkan Institute for Water and Environment (IB2E). https://www.researchgate.net/publication/348434478_The_Evaluation_of_Environmental_Situation_due_to_the_Oil_Industry_Activity_in_Albania

[19] Beqiraj, I., & Topi, D. (2016). Soil pollution from oil fields' exploitation in Albania-incidence of the marinza oil well explosion. *Mechanical engineering scientific journal (SKOPJE)*, *34*(1), 85–90. http://www.mesj.ukim.edu.mk/archive

[20] Métois, M., Benjelloun, M., Lasserre, C., Grandin, R., Barrier, L., Dushi, E., & Koçi, R. (2020). Subsidence associated with oil extraction, measured from time series analysis of Sentinel-1 data: Case study of the Patos-Marinza oil field, Albania. *Solid earth*, *11*(2), 363–378. https://doi.org/10.5194/se-11-363-2020

[21] Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical association*, *82*(397), 171-185. https://doi.org/10.1080/01621459.1987.10478410

[22] DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical science*, *11*(3), 189–228. https://doi.org/10.1214/ss/1032280214

[23] Efron, B., & Narasimhan, B. (2020). The automatic construction of bootstrap confidence intervals. *Journal of computational and graphical statistics*, *29*(3), 608–619. https://doi.org/10.1080/10618600.2020.1714633

[24] Steyerberg, E. W. (2019). *Clinical prediction models*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-16399-0

[25] Qiriazi, P., & Sala, S. (2000). Environmental problems of Albania. In *Remote sensing for environmental data in Albania: A strategy for integrated management* (pp. 13–30). Springer. https://doi.org/10.1007/978-94-011-4357-8_4

[26] National Environment Agency of Albania. (2019). *Report on the state of the environment 2019*. https://akm.gov.al/ova_doc/raport-per-gjendjen-e-mjedisit-2019/

[27] Gropa, O., Karamani, E., & Zema, R. (2025). Heavy oil production in the Patos-Marinza field: Mitigating high-viscosity constraints and enhancing recovery efficiency. *International journal of innovative research in science, engineering and technology*, *14*(7), 17090–17096. https://doi.org/10.15680/IJIRSET.2025.1407002

[28] Interstate Technology & Regulatory Council (ITRC). (2018). *TPH risk evaluation at petroleum-contaminated sites (TPHRisk-1)*. https://tphrisk-1.itrcweb.org

[29] Ditzler, C., Scheffe, K., & Monger, H. C. (2017). *Soil survey manual*. Government Printing Office. https://bibliotecadigital.ciren.cl/handle/20.500.13082/148329

[30] Programme., U. N. E. (2017). *Environmental assessment of Ogoniland*. https://www.unep.org/ogoniland

[31] Team, R. C. (2020). *RA language and environment for statistical computing, R foundation for statistical*. Computing. https://cir.nii.ac.jp/crid/1370298755636824325

[32] Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, *25*(24), 4279–4292. https://doi.org/10.1002/sim.2673

[33] Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

[34] Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into v. *BioScience*, *51*(5), 341–352. https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

[35] Millard, S. P. (2013). *EnvStats: An R package for environmental statistics*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-8456-1

[36] McCullagh, P. (2019). *Generalized linear models*. Routledge. https://doi.org/10.1007/978-1-4899-3242-6

[37] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International joint conference on artificial intelligence* (*IJCAI)* (Vol. 14, No. 2, pp. 1137-1145). IJCAI. https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf

[38] Hastie, T. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer. https://doi.org/10.1007/978-0-387-84858-7

[39] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[40] Molnar, C. (2020). *Interpretable machine learning*. Lulu. Com. https://christophm.github.io/interpretable-ml-book/

Appendix

Distribution and link

We model the positive response $Y_i>0$ using a Gamma distribution with a log link:

$$Y_i|\ x_i \sim \text{Gamma}(\mu_i, \phi),\ E[Y_i|\ x_i] = \mu_i,\ \text{Var}(Y_i|\ x_i) = \phi\mu_i^2,\ \log\mu_i = \eta_i,$$

with dispersion ϕ >0 and linear predictor η_i [36].

I. (GLM, Gamma–log).

$$\eta_i = \beta_0 + x_i^\top \beta,$$

where x_i contains distance, TOC, depth, moisture (on natural scales), and treatment contrasts for Soil Type and land use. Model parameters were obtained by maximum-likelihood estimation within the Gamma-family GLM framework [1], [36].

II. (GAM, Gamma–log).

$$\eta_i = \beta_0 + \sum_{j=1}^{J} s_j(x_{ij}) + z_i^\top \gamma,\ s_j(x) = \sum_{k=1}^{K_j} b_{jk}(x)\theta_{jk}\ ,$$

where $s_j(\cdot)$ are penalised cubic regression splines. Estimation proceeds by maximising a penalised likelihood with smoothing parameters λ_j typically estimated from the data (e.g., via REML) [4], [5]. At convergence of penalised Iteratively Reweighted Least Squares (IRLS), the model influence (smoother) matrix is defined as in [5], and effective degrees of freedom for each smooth are obtained from the corresponding partial influence matrix.

III. (SCAM, Gamma–log).

$$\eta_i = \beta_0 \sum_{j=1}^{J} s_j^{(cj)}(x_{ij}) + z_i^\top \gamma,$$

where each smooth has a spline basis expansion as in GAM

$$s_j^{(cj)}(x) = \sum_{k=1}^{K_j} b_{jk}(x)\theta_{kj},$$

but are estimated subject to shape constraints:

$$\max_\theta \{\ell(\theta) - \frac{1}{2}\sum_{j=1}^{J} \lambda_j \theta_j^\top S_j \theta_j\} \text{ subject to } C_j\theta_j \geq 0\ ,$$

with $\ell(\theta)$ the Gamma log-likelihood, S_j penalty matrices, λ_j smoothing parameters, and $C_j\theta_j \geq 0$ encoding monotonicity constraints. For monotonicity, constraints can be expressed as derivative sign restrictions on the smooths. Fitted shapes are verified by evaluating each smooth on a fine grid over the central 80% of covariate support and checking the sign of finite-difference derivatives [6].

Deviance and comparison metrics

Gamma unit deviance. The unit deviance is:

$$d(y_i,\mu_i)=2\left[\frac{y_i}{\mu_i}-\log\left(\frac{y_i}{\mu_i}\right)-1\right],$$

with total deviance $D=\sum_{i=1}^{n}d(y_i,\mu_i)$, and mean deviance $\bar{D}=\frac{1}{n}D$ [1], [36].

Error metrics.

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_i)^2},\quad \text{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\mu_i\right|.$$

Mean log-likelihood. Let $\ell(\hat{\theta})$ be the maximized log-likelihood and $\ell_i(\hat{\theta})$ the per-observation contribution, report $\bar{\ell}=\frac{1}{n}\ell(\theta)$.

AIC (ML).

$\text{AIC} = -2\,\ell(\theta) + 2k,$

where k is the number of estimated parameters.

Deviance explained. Using the null deviance D_{null},

$$\text{DE}=1-\frac{D}{D_{null}}.$$

Calibration (intercept/slope). We assessed calibration by regressing the observed log-response on the predicted linear predictor:

$$\log(Y_i)=a+b\hat{\eta}_i+\varepsilon_i$$

We then tested H_0: a=0, H_0: b=1, using standard OLS t-tests on the regression coefficients, with test statistics:

$$t_a=\frac{\hat{a}}{SE(\hat{a})},\ t_b=\frac{b-1}{SE(b)},$$

Well-calibrated models have a ≈ 0 and b ≈ 1 [24].

Local sensitivity (one-at-a-time, ±10%)

Let $\mu(x)$ be the predicted mean (response scale) from SCAM at a baseline X_0.

For covariate x_j, define symmetric ±10% perturbations with step h=0.10:

$$X_j^{(+)}:\ x_j^{(+)}=(1+h)x_{jo},\ X_j^{(-)}:\ x_j^{(-)}=(1-h)x_{jo},$$

holding all other covariates at numeric medians (factors at modal levels).

The Percent change in the mean is

$$\Delta(\%)=\frac{\mu(X_j^{(\pm)})-\mu(X_0)}{\mu(X_0)}\times 100.$$

A unit-free central elasticity (comparable across covariates) is:

$$E_j \sim \frac{\mu(X_j^{(+)}) - \mu(X_j^{(-)})}{2h\mu(X_0)},$$

which approximates the local log-derivative around X_0 [9], [10].

Bootstrap uncertainty for scenarios

For each scenario, we compute BCa 95% intervals from B bootstrap refits. Let $\hat{\theta}$ be the point estimate and $\{\hat{\theta}^*(b)\}_{b=1}^{B}$ the bootstrap replicates.

Bias-correction. Define the empirical Cumulative Distribution Function (CDF) at $\hat{\theta}$:

$$p = \frac{1}{B}\sum_{b=1}^{B} I\{\hat{\theta}^{*(b)} < \hat{\theta}\},$$

and set $z_0 = \Phi^{-1}(p)$ where Φ is the standard normal CDF.

Acceleration (jackknife). Using jackknife leave-one-out estimates $\{\hat{\theta}^{(-i)}\}_{i=1}^{n}$ and their mean $\bar{\theta}^{(-)}$,

$$a = \frac{\sum_{i=1}^{n}(\bar{\theta}^{(-)} - \hat{\theta}^{(-i)})^3}{6\left[\sum_{i=1}^{n}(\bar{\theta}^{(-)} - \hat{\theta}^{(-i)})^2\right]^{3/2}}.$$

Adjusted quantiles. With nominal tails α_L=0.025, α_U=0.975 and $z_\alpha=\Phi^{-1}(\alpha)$

$$\alpha_L^* = \Phi\left(z_0 + \frac{z_0 + z_{\alpha L}}{1 - a(z_0 + z_{\alpha L})}\right),\ \alpha_U^* = \Phi\left(z_0 + \frac{z_0 + z_{\alpha U}}{1 - a(z_0 + z_{\alpha U})}\right).$$

BCa interval. Let $\hat{\theta}^*_{(q)}$ denote the q-th empirical quantile of the bootstrap replicates (via the corresponding order statistic). The BCa 95% interval is

$$\left[\hat{\theta}^*_{(\alpha_L^*)}, \hat{\theta}^*_{(\alpha_U^*)}\right].$$

In this study, we set B=5,000, and follow standard BCa construction [21–23].