Received: 30 March 2022 | Revised: 25 October 2022 | Accepted: 24 March 2023

DOI: 10.1111/stan.12293

ORIGINAL ARTICLE

WILEY

A robust mixed-effects parametric quantile regression model for continuous proportions: Quantifying the constraints to vitality in cushion plants

Divan A. Burger[1,2] | **Sean van der Merwe**[3] | **Emmanuel Lesaffre**[4] | **Peter C. le Roux**[5] | **Morgan J. Raath-Krüger**[5]

[1]Cytel Inc., 1050 Winter Street Waltham, 02451, MA, USA

[2]Department of Statistics, University of Pretoria, Pretoria, South Africa

[3]Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, South Africa

[4]I-BioStat, KU Leuven, Leuven, Belgium

[5]Department of Plant and Soil Sciences, University of Pretoria, Pretoria, South Africa

Correspondence
Divan A. Burger, Department of Statistics, University of Pretoria, Pretoria 0028, South Africa.
Email: divanaburger@gmail.com

Funding information
NRF of South Africa (Grant Number 132383 & Postdoctoral Grant PDG190329424983); South African National Antarctic Programme (Grant Numbers 93077 & 110726)

There is no literature on outlier-robust parametric mixed-effects quantile regression models for continuous proportion data as an alternative to systematically identifying and eliminating outliers. To fill this gap, we formulate a robust method by extending the recently proposed fixed-effects quantile regression model based on the heavy-tailed Johnson-t distribution for continuous proportion data to the mixed-effects modeling context, using a Bayesian approach. Our proposed method is motivated by and used to model the extreme quantiles of the vitality of cushion plants to provide insights into the ecology of the system in which the plants are dominant. We conducted a simulation study to assess the new method's performance and robustness to outliers. We show that the new model has good accuracy and confidence interval coverage properties and is remarkably robust to outliers. In contrast, our study demonstrates that the current approach in the literature for modeling hierarchically structured bounded data's quantiles is susceptible to outliers, especially when modeling the extreme quantiles. We conclude that the proposed model is an appropriate robust alternative to the cur-

rent approach for modeling the quantiles of correlated continuous proportions when outliers are present in the data.

KEYWORDS

Bayesian, continuous proportions, cushion plants, mixed-effects, outliers, quantile regression, sub-Antarctic

1 | MODELING CONTINUOUS PROPORTIONS

1.1 | Need for robust quantile regression models

Traditional regression approaches focus on modeling the relationship between the average response and a set of covariates. As an alternative to modeling the mean response, regression curves can be fitted to selected parts of the response variable's distribution through quantile regression (Cade & Noon, 2003; Koenker & Bassett Jr, 1978), thereby modeling the relationship between the outcome variable and covariates for any portion (e.g., *extreme quantiles*) of the probability distribution. For example, Wei, Kehm, Goldberg, and Terry (2019)'s study showed that certain risk factors have a limited impact on adult BMI's lower quantiles but are significantly associated with the mean BMI. Thus, quantile regression allows quantifying the strength of the association between the response and the set of covariates that may go unnoticed when only the mean response is considered, as conventional techniques do. Quantile regression can be performed using either parametric or nonparametric methods (Min & Kim, 2004; Yirga, Melesse, Mwambi, & Ayele, 2021).

This paper focuses on modeling the quantiles of correlated data bound to the unit interval using random effects. The proposed model is applied to an exemplar dataset from an ecological study. Ideally, one would consider using nonparametric quantile regression methods that do not make any distributional assumptions about the response variable or random effects. However, no nonparametric methods for implementing random-effects models for bounded outcomes are available in the literature. We choose a fully parametric Bayesian approach for the current paper by implementing the models in `JAGS` (Plummer, 2003). Unlike nonparametric quantile regression models, a significant disadvantage of parametric quantile regression models is that the inference about their parameters can be sensitive to outliers. Our application dataset contains significant outliers; therefore, we particularly focus on models that are robust to outliers. As software, we choose `JAGS` because it is convenient, user-friendly, relatively fast, and can handle complex models with many parameters (such as random-effects models). Finally, the choice of the Bayesian approach is motivated by the fact that Bayesian inference does not rely on asymptotic theory as frequentist methods generally do, and therefore Bayesian confidence intervals often show better coverage than their frequentist counterparts.

1.2 | Models for mean and quantiles in the literature

Bounded responses can be modeled using beta regression when the mean response is of interest (Ferrari & Cribari-Neto, 2004) and the Kumaraswamy, unit-Weibull, and unit-Birnbaum–Saunders models (Mazucheli, Leiva, Alves, & Menezes, 2021; Mazucheli,

Menezes, Fernandes, de Oliveira, & Ghitany, 2020; Mitnik & Baek, 2013) when the responses' quantiles are of interest.

Ecological and clinical research commonly encounters extreme or unusual observations in the data, that is, "outliers" (Benhadi-Marín, 2018). An *outlier* is a data point that differs considerably from other data points. Such an outlier is called influential if the important features of typical analyses would be altered if they were to be retained or deleted from the dataset (Begashaw & Yohannes, 2020). Several different mechanisms can result in outliers in the data, such as sampling errors (Begashaw & Yohannes, 2020). In contrast, natural outliers that are not due to a sampling error may also emerge. There has been considerable discussion about handling outliers in a variety of research fields (Benhadi-Marín, 2018; Kwak & Kim, 2017; Leys, Klein, Dominicy, & Ley, 2018). One commonly implemented technique is the systematic identification and elimination of outliers. However, deleting outliers may fail to compensate for the uncertainty in the exclusion process and, consequently, may result in underestimated SEs of estimates (Lange, Little, & Taylor, 1989). An alternative to excluding outliers from the data is to employ robust regression techniques that downweigh the influence of outliers on statistical inference.

In the context of parametric regression modeling, one way to achieve robustness to outliers is by assuming heavy-tailed distributions for response variables. For example, replacing the conventional normal distribution for the response with the t-distribution yields inference about the mean outcome robust to outliers (Lange et al., 1989). Alternatively, the modeling of the mean response can be substituted by that of the median, which is considered a more robust measure of central tendency when the data exhibit skewness and contain outliers (Burger & Lesaffre, 2021). However, in parametric quantile regression (Burger & Lesaffre, 2021; Cancho, Bazán, & Dey, 2020), the estimates of the regression coefficients, even for the median, can be prone to outliers if the underlying distribution does not accommodate skewness or heavy tails. Therefore, in the presence of outliers, heavy-tailed distributions can be considered for robust parametric quantile regression modeling, similar to robustly regressing the mean.

For the robust modeling of the average of continuous proportion data, the rectangular beta and flexible beta regression models were proposed by Bayes, Bazán, and García (2012) and Migliorati, di Brisco, & Ongaro, 2018, respectively, to replace the conventional beta model (when outliers are present). di Brisco and Migliorati (2020), for example, modeled EQ-VAS scores, a patient-reported outcome ranging from 0% to 100%, in a Parkinson's disease (PD) longitudinal study. Since the PD dataset of di Brisco and Migliorati (2020) contains outliers (e.g., due to a small group of outliers in EQ-VAS scores near 0), they used the mixed-effects augmented flexible beta model to model the mean EQ-VAS scores over time robustly.

For robust modeling of the quantiles of bounded data, the recently proposed heavy-tailed unit-interval distributions, namely the power normal-logistic distribution (Cancho et al., 2020) and the Johnson-t distribution (Lemonte & Moreno-Arenas, 2020), can serve as alternatives to the Kumaraswamy distribution. The robust logistic quantile regression model of Galarza, Zhang, and Lachos (2020) can also model the quantiles of bounded data in the presence of outliers.

1.3 | Objectives and outline

To the best of our knowledge, the Kumaraswamy mixed model of Bayes, Bazán, and de Castro (2017) is the only parametric mixed-effects quantile regression approach available in the literature for bounded data. The Kumaraswamy distribution is available in `R-INLA` (Lindgren & Rue, 2015) and can be implemented as a random-effects model; Flores, Prates, Bazán, and

Bolfarine (2021) applied it in a spatial modeling context. However, the Kumaraswamy distribution consists of two parameters and lacks flexibility regarding its tails (mainly, it cannot accommodate heavy tails). Hence, estimates of the regression coefficients from the Kumaraswamy model may be prone to outliers in the data. We note that there is no literature on parametric mixed-effects quantile regression models for bounded data that can accommodate outliers.

This paper proposes a parametric mixed-effects quantile regression model for bounded outcomes that is robust to outliers in the data. We extend the recently proposed fixed-effects quantile regression model based on the Johnson-*t* distribution of Lemonte and Moreno-Arenas (2020) for bounded (continuous proportion) outcomes to the mixed-effects modeling context. Hence, we provide a robust quantile regression method for hierarchically structured data using a mixed-effects approach. We compare the mixed-effects Johnson-*t* and Kumaraswamy models to assess the suitability of our model. That is, we contrast the current model, namely, the Kumaraswamy model, for hierarchically structured bounded data, with our proposed robust method, namely, the Johnson-*t* model. We also consider robust and nonrobust models for the mean outcome, respectively, based on the conventional and rectangular beta distributions. We limit our study to the rectangular beta model for robustly modeling the mean (alternatives include the model of Migliorati et al., 2018 (mentioned earlier)).

The paper is organized as follows: Section 2 describes the ecological dataset that motivates our proposed methodology. Section 3 introduces the nonrobust and robust Bayesian mixed-effects models for bounded outcome data, namely the beta, rectangular beta, Kumaraswamy, and Johnson-*t* models. Section 4 applies the mixed-effects models to the ecological dataset. Section 5 presents simulation studies to investigate the robustness of the Kumaraswamy and Johnson-*t* models to outliers (data contamination), as well as assess the performance of the Johnson-*t* model. Finally, Section 6 presents a discussion of the results and findings of the paper.

2 | ECOLOGY STUDY

2.1 | Cushion plant vitality

Raath-Krüger et al. (2022) compiled a long-term dataset of repeated measures to examine the impact of the grass species *Agrostis magellanica* on the cushion-forming plant, *Azorella selago*, using sub-Antarctic Marion Island as a model system. These two species are the dominant vascular plants in the sub-Antarctic, and *A. magellanica* is the most common vascular plant species growing on *A. selago* (Huntley, 1972). Because of its cushion growth form, *A. selago* can modify the local microenvironment (e.g., ameliorate temperature conditions (McGeoch, le Roux, Hugo, & Nyakatya, 2008; Nyakatya & McGeoch, 2008)) and therefore has the potential to positively impact species associated with it, particularly in cold, wind-exposed areas where the cushion plant is commonly found. Specifically, *A. selago* is known to have a strong positive impact on the cover, reproductive output, and abundance of *A. magellanica*, compared to surrounding areas where *A. selago* is absent (Raath-Krüger, Schöb, McGeoch, & le Roux, 2021). However, little is known about the reciprocal impact of *A. magellanica* on *A. selago*. Therefore, in their study, Raath-Krüger et al. (2022) documented the long-term outcome of the *Azorella-Agrostis* interaction by assessing changes in *A. selago* vitality (i.e., dead stem cover) in relation to *A. magellanica* cover over a 13-year time period. In the present study, we analyze a subset of the data compiled by Raath-Krüger et al. (2022), specifically examining the effect of *A. magellanica* cover, the cover of other vascular plants and mosses, altitude (high vs. mid), and aspect (west vs. east) on *A. selago*

vitality between 2003 (initial year of the survey) and 2016 (final year of the survey). We consider *A. selago* dead stem cover as a proxy for cushion plant vitality (see le Roux, McGeoch, Nyakatya, & Chown, 2005).

2.2 | Objectives

Our proposed statistical method is motivated by modeling the extreme quantiles of the vitality (i.e., dead stem cover) of *A. selago*, which may provide several key insights into the ecology of the system in which *A. selago* cushion plants are dominant. The ecological research questions we aim to address are as follows:

1. Given that *A. selago* vitality could be negatively affected by the cover of *A. magellanica* and other vascular and nonvascular plant species (see le Roux et al., 2005; Owen, 1995; Raath-Krüger et al., 2022), we ask: *Is the limit to* `A. selago` *vitality constrained by the cover of vascular and nonvascular plant species growing on* `A. selago`*?* We model the 0.95th quantile of dead stem cover to address this research question. Here, the 0.95th quantile quantifies the upper extreme quantile of cushion plant vitality, representing the unhealthiest cushion plants.
2. Second, given that (i) the western and eastern aspects of Marion island are abiotically different (see, e.g., Goddard, Craig, Schoombie, & le Roux, 2022), resulting in plant populations across the two aspects experiencing different abiotic stressors and (ii) plant species on Marion Island are exposed to increasingly stressful abiotic conditions with increasing altitude (see le Roux, 2008), we ask: *Is the limit to* `A. selago` *vitality imposed by altitude and aspect?* Similar to the above, we model the 0.95th quantile of dead stem cover to address this research question.
3. Given that environmental conditions may have different effects on the upper or lower limits of an organism's health (broadly in line with, for example, Wei et al., 2019), we ask: *Does the effect of vascular and nonvascular plant species, altitude, and aspect on* `A. selago` *vary across plants of different vitality?* As a result, we model a broad range of quantiles of dead stem cover, particularly the 0.1th, 0.25th, 0.5th, 0.75th, and 0.95th quantiles, to address this research question. Here, the 0.1th quantile quantifies the vitality of the cushion plants' lower extreme quantile, representing the healthiest cushion plants.

We also model the mean dead stem cover to compare the median and mean fits for completeness sake. Since bounded data tends to be skewed, the median may be a better measure of central tendency than the mean.

2.3 | Model covariates

We test the long-term constraints on cushion plant vitality by modeling the mean and the quantiles of the final *A. selago* dead stem cover in response to (i) initial *A. selago* size, (ii) initial *A. magellanica* cover on *A. selago*, (iii) initial combined cover of other vascular plants and mosses on *A. selago*, (iv) altitude, and (v) aspect. Note that the dead stem cover on *A. selago* denotes the proportion of black and grey parts on the cushion plant and is therefore considered a continuous proportion outcome. We also include the initial *A. selago* dead stem cover in the model as a predictor variable because we expect *A. selago* individuals with greater dead stem cover in the initial year of the survey to have increasingly more dead stem cover in the final year. In order

to account for the data's spatial structure, we include a random effect for "plot" in the model. The plots represent specific sampling sites situated on the eastern and western aspects of Marion Island at mid and high altitudes, from which *A. selago* individuals were surveyed. The data from the low-altitude sites of Raath-Krüger et al. (2022) have been excluded since the low and mid altitudes sites exhibited similar patterns.

We specify the covariates for modeling the continuous proportion of dead stem cover on *Azorella* individual *j* of plot *i* measured in 2016 as follows:

- DS_{ij}, Azo_{ij}, Agr_{ij}, and CO_{ij} denote the covariates corresponding to *Azorella* individual *j* of plot *i* taken in 2003 (i.e., initial measurements): DS_{ij} is the *Azorella* dead stem cover (%), Azo_{ij} the *Azorella* size (cm^2; log-transformed), Agr_{ij} the *Agrostis* cover on *Azorella* (%), and CO_{ij} is the combined cover of other vascular plants and mosses on *Azorella* (%).
- The indicator variables Mid_i and $West_i$ are assigned as follows: $Mid_i = 1$ if plot *i* is situated at a mid-altitude, and $Mid_i = 0$ otherwise (i.e., high altitudes); $West_i = 1$ if plot *i* is situated on the western side of the island, and $West_i = 0$ otherwise (i.e., eastern side).

The linear predictor corresponding to the final dead stem cover on *Azorella* individual *j* of plot *i* is written as:

$$\beta_0 + \beta_1 DS_{ij} + \beta_2 Azo_{ij} + \beta_3 Agr_{ij} + \beta_4 CO_{ij} + \beta_5 Mid_i + \beta_6 West_i + u_i, \quad (1)$$

where $\beta_0, \beta_1, \ldots, \beta_6$ are the corresponding regression coefficients (fixed effects), and u_i is the random intercept of plot *i*.

2.4 | The dataset and outliers

In their original analysis, Raath-Krüger et al. (2022) addressed one of their particular research questions by modeling the mean final dead stem cover in response to the covariates above. The data analysis considered the fit of a beta regression model to the final dead stem cover based on the logit link function (i.e., modeling the mean outcome) using the `R` package `glmmTMB` (Brooks et al., 2017). Raath-Krüger et al. (2022) excluded *A. selago* individuals that died (i.e., data points considered as significant outliers due to catastrophic loss of biomass) over the 13-year observation period from their analysis. In contrast, in the present paper, we opt for drawing inferences about the quantiles of dead stem cover for this dataset while accommodating and safeguarding against outliers by employing robust techniques without excluding specific data points (outliers) from the analysis.

Figure 1 shows the final versus initial *A. selago* dead stem cover by plot, altitude, and aspect, distinguishing between the data included and excluded from the original analysis. For certain plots, it is evident that outliers (relative to the relationship between the final vs. initial dead stem cover) are primarily associated with the previously excluded data. Aside from the previously excluded data due to catastrophic loss, we note that dead stem cover on a few plants (included in the previous analysis of Raath-Krüger et al., 2022) changed considerably during the monitoring period, thus also resulting in data outliers. Figure S1a,b in Section A of Data S1 gives a photograph example of the dead stem cover that remained similar on the vast majority of individual plants over the 13-year monitoring period, not resulting in an outlier. In contrast, Figures S1c,d is an example of an individual plant that yielded an outlier; in this case, the

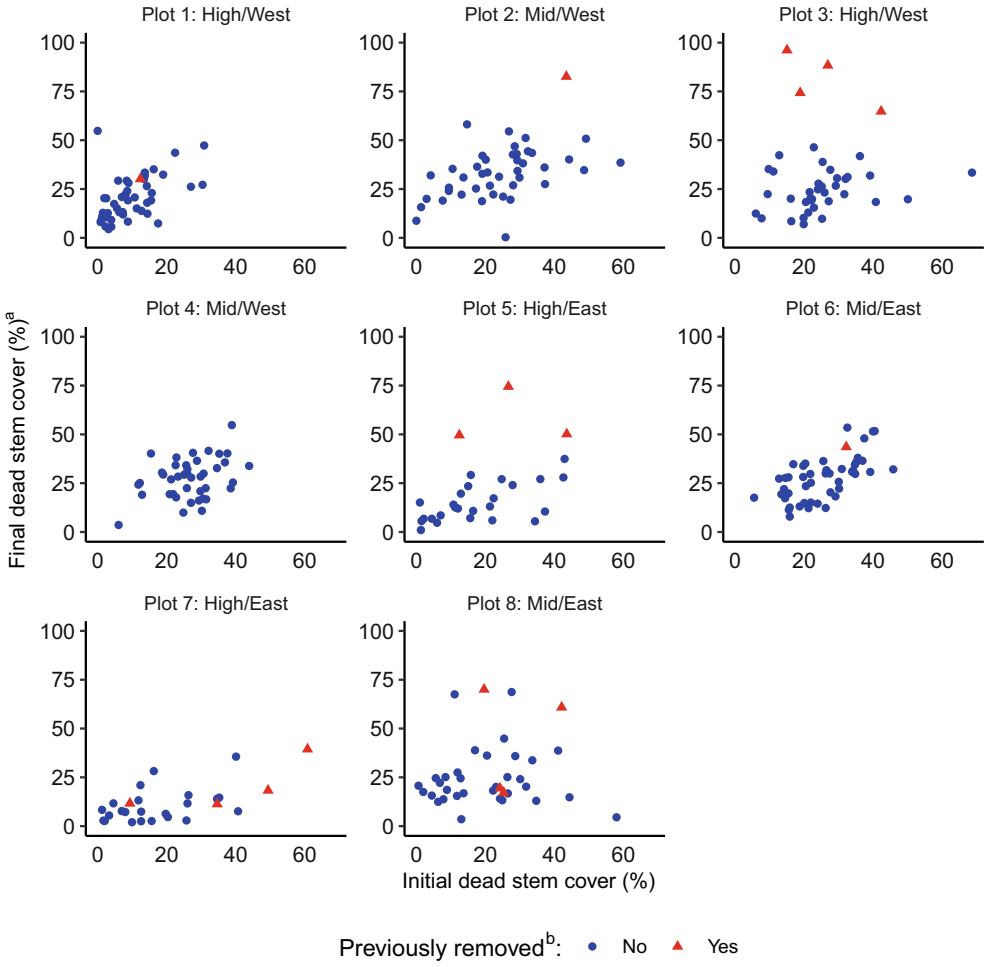

FIGURE 1 Cushion plant dataset: final versus initial *Azorella selago* dead stem cover by plot, altitude, and aspect. [a]Plots are situated either at a mid or high altitude on the western or eastern side of the island. [b]The red triangles represent the data originally excluded from Raath-Krüger et al. (2022)'s analysis; in contrast, the blue dots represent the data originally included in the study of Raath-Krüger et al. (2022)

dead stem cover on the individual plant increased from about 1% to 50% during the monitoring period.

The dataset we consider consists of eight plots and 308 cushion plants for the current publication, including the previously excluded data points of the original analysis. Hence, we model the outliers in the current paper instead of excluding them.

2.5 | Descriptive analysis

Summary statistics of the final dead stem cover are presented in Table 1. The quantiles of the final dead stem cover of the island's western side are higher than those of the eastern side, and the quantiles at mid-altitudes are higher than those at high altitudes. The preliminary investigation of the data, particularly the 0.75th and 0.95th quantiles, suggests that the limit to the cushion

TABLE 1 Cushion plant dataset: descriptive statistics of final *Azorella selago* dead stem cover by variable.

Variable [a]		Mean	SD	Quantile				
				10%	25%	50%	75%	95%
Altitude	Mid	28.8	13.44	13.8	19.2	27.8	36.0	53.5
	High	21.3	16.34	5.7	10.1	18.3	28.1	49.9
Aspect	West	27.8	15.07	10.2	18.7	26.5	34.7	54.5
	East	22.6	15.06	5.6	12.2	19.4	30.7	51.7

Abbreviations: SD, standard deviation.

[a]*A. selago* individuals were surveyed at sampling sites situated on the eastern and western aspects of Marion Island at mid and high altitudes.

plants' vitality may be imposed by altitude and aspect. However, a regression analysis considering the covariates, as mentioned earlier, needs to be performed to make formal conclusions about the constraints to the vitality of cushion plants in this system.

3 | MIXED-EFFECTS MODELS FOR CONTINUOUS PROPORTIONS

This section formulates the robust mixed-effects regression model for the quantiles of bounded responses, namely the Johnson-t model and its nonrobust competitor, the Kumaraswamy model of Bayes et al. (2017). In addition, we consider the rectangular beta model of Bayes et al. (2012) and its nonrobust competitor, namely the conventional beta model of Ferrari and Cribari-Neto (2004). The latter two models fit the mean as a function of the covariates, and we are interested in comparing the results of these models with the results of the two quantile regression models.

For the four models under consideration in this paper (namely beta, rectangular beta, Kumaraswamy, and Johnson-t; see Sections 3.1–3.4), we use the logit link function to model the mean and quantiles as a function of covariates.

Suppose that y_{ij} is the bounded outcome for cluster $i = 1, \ldots, I$ and observation $j = 1, \ldots, J_i$. Let $\boldsymbol{\beta}$ and $\boldsymbol{u_i}$ denote fixed and cluster-specific random effects vectors, and $\boldsymbol{x_{ij}}$ and $\boldsymbol{z_{ij}}$ the covariate vectors, respectively. Assume the $\boldsymbol{u_i}$ follow a multivariate normal distribution with mean $\mathbf{0}$ and d-dimensional unstructured covariance matrix $\boldsymbol{\Sigma}$, such that $\boldsymbol{u_i} \,|\boldsymbol{\Sigma} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$.

Details on the Bayesian specification of the candidate models are presented in Section 3.5.

3.1 | Beta regression model

The probability density function of the reparameterized beta regression model of Ferrari and Cribari-Neto (2004) for a given bounded outcome $0 < y_{ij} < 1$ is:

$$f\left(y_{ij} \,|\boldsymbol{\beta}, \boldsymbol{u}_i, \rho\right) = \frac{\Gamma(\rho)}{\Gamma\left(\kappa_{ij}\rho\right)\Gamma\left(\rho\left(1-\kappa_{ij}\right)\right)} y_{ij}^{\kappa_{ij}\rho-1}\left(1-y_{ij}\right)^{\rho(1-\kappa_{ij})-1},$$

where $\kappa_{ij} = \frac{\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\boldsymbol{z}_{ij}'\boldsymbol{u}_i\right)}{1+\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\boldsymbol{z}_{ij}'\boldsymbol{u}_i\right)}$ is the conditional mean of y_{ij} given the ith random effect under the beta model, and $\rho > 0$ is the precision parameter of the beta distribution.

It is noted that the beta distribution does not accommodate heavy-tailed data (Bayes et al., 2012), and therefore, the estimation of the mean parameter is prone to data outliers. Moreover, the beta distribution's quantile function is not available in closed form, making the beta distribution not appropriate for parametric quantile regression modeling.

3.2 | Rectangular beta regression model

The probability density function of the rectangular beta regression model of Bayes et al. (2012) for a given bounded outcome $0 < y_{ij} < 1$ is:

$$f\left(y_{ij}\,|\boldsymbol{\beta}, \boldsymbol{u}_i, \rho, \phi\right) = \phi\left(1 - \left|2\kappa_{ij} - 1\right|\right) + \left[1 - \phi\left(1 - \left|2\kappa_{ij} - 1\right|\right)\right] f_b \times \left(y_{ij}\left|\frac{\kappa_{ij} - 0.5\phi\left(1 - \left|2\kappa_{ij} - 1\right|\right)}{1 - \phi\left(1 - \left|2\kappa_{ij} - 1\right|\right)}, \rho\right),$$

where $\kappa_{ij} = \frac{\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{u}_i\right)}{1+\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{u}_i\right)}$ is the conditional mean of y_{ij} given the ith random effect under the rectangular beta model, $f_b\left(x\,|a_1, a_2\right) = \frac{\Gamma(a_2)}{\Gamma(a_1 a_2)\Gamma(a_2(1-a_1))} x^{a_1 a_2 - 1}(1-x)^{a_2(1-a_1)-1}$, and $\rho > 0$ and $0 \leq \phi \leq 1$ are, respectively, the precision and shape parameters of the rectangular beta distribution.

The parameter ϕ governs the tail of the rectangular beta distribution: larger values of ϕ yield heavier tails, making the distribution more robust to outliers than the conventional beta distribution (Bayes et al., 2012). The probability density function of the rectangular beta model reduces to that of the conventional beta model when $\phi = 0$.

Even though the rectangular beta model accommodates heavy-tailed data and gross outliers, its quantile function cannot be expressed analytically. Therefore, this model is not suitable for quantile regression modeling.

3.3 | Kumaraswamy regression model

The probability density function of the Kumaraswamy regression model of Bayes et al. (2017) for a given bounded outcome $0 < y_{ij} < 1$ is:

$$f\left(y_{ij}\,|\boldsymbol{\beta}, \boldsymbol{u}_i, \rho\right) = -\frac{\log\left(1-q\right)\rho}{\log\left(1-e^{-\rho}\right)\log\left(\kappa_{ij}\right)} y_{ij}^{-\frac{\rho}{\log\left(\kappa_{ij}\right)}-1}\left(1 - y_{ij}^{-\frac{\rho}{\log\left(\kappa_{ij}\right)}}\right)^{\frac{\log(1-q)}{\log\left(1-e^{-\rho}\right)}-1},$$

where $\kappa_{ij} = \frac{\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{u}_i\right)}{1+\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{u}_i\right)}$ is the qth conditional quantile of y_{ij} given the ith random effect under the Kumaraswamy model, and $\rho > 0$ is the precision parameter of the Kumaraswamy distribution.

The Kumaraswamy model is suitable for quantile regression because the quantile function of the Kumaraswamy distribution can be expressed analytically. However, similar to the beta distribution, the Kumaraswamy distribution has only a location and precision parameter but no shape parameter, as the rectangular beta distribution does. Hence, the Kumaraswamy model may be prone to outliers in the data.

3.4 | Johnson-*t* regression model

The probability density function of the Johnson-*t* regression model of Lemonte and Moreno-Arenas (2020) for a given bounded outcome $0 < y_{ij} < 1$ is:

$$f\left(y_{ij}\,|\boldsymbol{\beta}, \boldsymbol{u}_i, \rho, \nu\right) = \frac{\rho\nu^{\frac{1}{2}\nu}}{y_{ij}\left(1-y_{ij}\right)} \frac{\Gamma\left(\frac{1}{2}+\frac{1}{2}\nu\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\nu\right)} \times \left[\nu + \left\{t_\nu\left(q\right) + \rho\left[\log\left(\frac{y_{ij}}{1-y_{ij}}\right) - \log\left(\frac{\kappa_{ij}}{1-\kappa_{ij}}\right)\right]\right\}^2\right]^{-\frac{\nu+1}{2}},$$

where $\kappa_{ij} = \frac{\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\boldsymbol{z}_{ij}'\boldsymbol{u}_i\right)}{1+\exp\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta}+\boldsymbol{z}_{ij}'\boldsymbol{u}_i\right)}$ is the qth conditional quantile of y_{ij} given the ith random effect under the Johnson-*t* model, $\rho > 0$ and $\nu > 0$ are respectively the dispersion parameter and degrees of freedom of the Johnson-*t* distribution, and $t_\nu(q)$ is the qth quantile of the conventional *t*-distribution with degrees of freedom ν.

Figure S2 in Section A of Data S1 shows examples of the Johnson-*t* distribution's probability density function for various values of q and ν. The parameter ν governs the tail of the Johnson-*t* distribution: smaller values of ν yield heavier tails (Lemonte & Moreno-Arenas, 2020). The Johnson-*t* distribution reduces to the Johnson-normal distribution (Johnson, 1949) for infinite degrees of freedom ($\nu \to \infty$). Hence, the Johnson-*t* distribution accommodates heavy-tailed data (i.e., outliers).

The Johnson-*t* distribution accommodates heavy-tailed data (i.e., outliers) and is suitable for quantile regression given its closed-form quantile function.

3.5 | Bayesian specification

The prior distributions are specified in such a way as to assure vagueness about prior belief on the model parameters (i.e., *weakly* informative priors).

A normal prior distribution, namely Normal $(0, 10{,}000)$, is specified for each component of the vector of fixed effects (i.e., $\boldsymbol{\beta}$) for each regression model.

The precision parameter of the beta, rectangular beta, and Kumaraswamy models (i.e., ρ) and the dispersion parameter of the Johnson-*t* model (i.e., ρ) are assigned a gamma prior distribution, namely Gamma $(0.0001, 0.0001)$.

The exponential distribution is often used as a prior distribution for the degrees of freedom of the *t*-distribution. However, the exponential distribution has a relatively light tail, and therefore, the degrees of freedom's posterior distribution may be unduly influenced by the exponential prior distribution, in some cases, yielding confidence interval coverage very far below the nominal value (Simpson, Rue, Riebler, Martins, & Sørbye, 2017). Therefore, as an alternative to the widely used exponential prior distribution, the Johnson-*t* distribution's degrees of freedom (i.e., ν) are assigned the hierarchical prior distribution of Juárez and Steel (2010), which is more heavy-tailed relative to the exponential and gamma distributions. In particular, the hierarchical prior distribution of ν is expressed as a mixture of an exponential distribution with rate parameter 1 for ϵ, namely *Exp* (1), and a gamma distribution with shape parameter 2 and rate parameter ϵ, namely

Gamma $(2, \epsilon)$. From the law of total probability (integrating out ϵ), the resulting probability density function of ν is written as $P(\nu) = \frac{2\nu}{(1+\nu)^3}$.

The shape parameter of the rectangular beta distribution (i.e., ϕ) is assigned a uniform prior distribution, namely Uniform $(0, 1)$.

We specify the matrix-generalized half-t (MGH-t) prior distribution of Huang and Wand (2013) for the variance-covariance matrix (i.e., $\mathbf{\Sigma}$) for each model as a more appropriate alternative to the conventional inverse Wishart distribution. The MGH-t prior distribution of $\mathbf{\Sigma}$ is expressed as a mixture representation of Gamma $(0.5, 0.25)$ for the diagonal entries of diagonal matrix $\mathbf{\Omega} = \text{diag}(\omega_1, \ldots, \omega_z, \ldots, \omega_d)$, and a Wishart distribution with inverse scale matrix $4\mathbf{\Omega}$ and degrees of freedom $d + 1$, namely Wishart $(4\mathbf{\Omega}, d + 1)$ (Burger, Schall, Ferreira, & Chen, 2020). This mixture representation results in the specification of the half-t prior distribution with location parameter 0, scale parameter 4, and two degrees of freedom, namely $t(0, 4, 2)\, T(0, \infty)$, for the SD terms in $\mathbf{\Sigma}$, and the uniform prior distribution, namely Uniform $(-1, 1)$, for the correlation terms in $\mathbf{\Sigma}$. From the law of total probability, the set of nuisance parameters $\mathbf{\Omega}$ integrated out results in the MGH-t prior distribution, namely:

$$P(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-d-1} \prod_{z=1}^{d} \left[2 \left(\mathbf{\Sigma^{-1}} \right)_{zz} + 0.25 \right]^{-\frac{d+2}{2}},$$

where $\mathbf{\Sigma} > \mathbf{0}$, and $\left(\mathbf{\Sigma^{-1}}\right)_{zz}$ is the zth diagonal entry of $\mathbf{\Sigma^{-1}}$.

The MCMC Gibbs sampling algorithm can draw samples from the joint posterior distribution of the model parameters obtained by forming the product of all likelihoods and prior distributions (Gelfand & Smith, 1990). Software such as JAGS (Plummer, 2003) can be employed to carry out the Gibbs sampling procedure.

For each model, 150,000 samples were simulated from the joint posterior distribution for 15 parallel chains. Among those 150,000 samples (per chain), the initial 10,000 samples were discarded (burn-in). The convergence of posterior samples was checked using trace plots and Brooks-Gelman-Rubin statistics (Brooks & Gelman, 1998). We used a thinning factor of 25 to reduce autocorrelation among the samples. Ultimately, we obtained 84,000 posterior samples in total for each model parameter (hence, $K = 84{,}000$).

We reported the posterior distributions' mean and highest posterior density (HPD) intervals as point and interval estimates of the model parameters. The 95% HPD intervals are constructed by finding the shortest interval covering 95% of posterior samples and thus are more informative than the classic symmetric interval.

4 | DATA ANALYSIS

This section examines the fit of the four regression models to the cushion plant data. Comparing the fit of the Kumaraswamy model with that of the Johnson-t model illustrates the impact of outliers on the two model fits. Furthermore, a comparison of the parameter estimates of the beta and rectangular beta distribution, on the one hand, and the Kumaraswamy and Johnson-t distribution, on the other hand, illustrates the difference in modeling the mean compared to the quantiles of the bounded data. We then discriminate between models formally, assess model adequacy, and address the ecological research questions based on the results from the *preferred* mean and quantile models.

4.1 | Model implementation

We report regression fits of the mean and the $q \in \{0.1, 0.25, 0.5, 0.75, 0.95\}$ quantiles of dead stem cover to answer the ecological research questions outlined in Section 2. The extreme quantiles of the cushion plant vitality quantified by the 0.1th and 0.95th quantiles of dead stem cover are of primary interest.

We fitted the mixed-effects beta, rectangular beta, Kumaraswamy, and Johnson-t regression models in Section 3 to the proportion of dead stem cover in 2016 (see Section 2). We are primarily interested in evaluating the robustness of the Johnson-t and Kumaraswamy models to the outliers present in our dataset.

As per Equation (1), the terms in the linear predictor $\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{u}_i$ (Section 3) are defined as follows: $\boldsymbol{x}_{ij} = \left(1, \mathrm{DS}_{ij}, \mathrm{Azo}_{ij}, \mathrm{Agr}_{ij}, \mathrm{CO}_{ij}, \mathrm{Mid}_i, \mathrm{West}_i\right)'$ and $\boldsymbol{z}_{ij} = 1$ are the covariate vectors, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ \ldots \ , \beta_6)'$ and $\boldsymbol{u}_i = u_i$ are respectively the vectors of fixed and random effects. Furthermore, $\boldsymbol{\Sigma}$ is the unstructured covariance matrix of $\boldsymbol{u}_i$. Since the model contains a single random effect, its variance component is expressed in scalar rather than vector form (i.e., $\boldsymbol{\Sigma} = \sigma^2$).

The models were fitted using `JAGS` (Plummer, 2003) via the `R` package `jagsUI` (Kellner, 2021). We ran our models on a desktop computer with a 3.00 GHz Intel® CoreTM i9-10980XE processor and 64 GB installed memory (RAM).

4.2 | Regression fits and model comparison

This section presents the mean and quantile regression fits derived from the candidate regression models. In addition, we calculated the deviance information criterion (DIC) statistic marginalized over the models' random effects to compare the candidate models, that is, the marginal DIC (mDIC) of Quintero and Lesaffre (2018). The mDIC statistic is a more appropriate model comparison tool than the conventional DIC statistic of Spiegelhalter, Best, Carlin, and van der Linde (2002) (i.e., conditional on the random effects) when population-average inferences are of interest, which is the case for our application. Details on the calculation of the mDIC are presented in Appendix A.

We first compare the fits of the nonrobust and robust quantile models (i.e., the Kumaraswamy and Johnson-t models). We do this by studying the posterior estimates (PEs) and 95% highest posterior density (HPD) intervals for the regression coefficients. The quantile regression coefficients' PEs and 95% HPD intervals calculated from the Kumaraswamy and Johnson-t models are presented in Figures 2 and S3 in Section A of Data S1. In addition, the complete set of quantile model parameters, including the model comparison statistic, are presented in Table S1 in Section B of Data S1. From Figures 2 and S3, and Table S1, we observe the following:

- The mDIC strongly favors the robust quantile model over the nonrobust model.
- The degrees of freedom estimate under the Johnson-t model confirms our finding that the data are heavy-tailed ($\hat{\nu} \approx 3$), implying that the robust model better fits the data than the nonrobust model.
- The regression coefficients' 95% HPD intervals under the nonrobust quantile model are generally wider than those under the robust model. Therefore, the robust model produces shorter 95% HPD intervals for the regression coefficients.

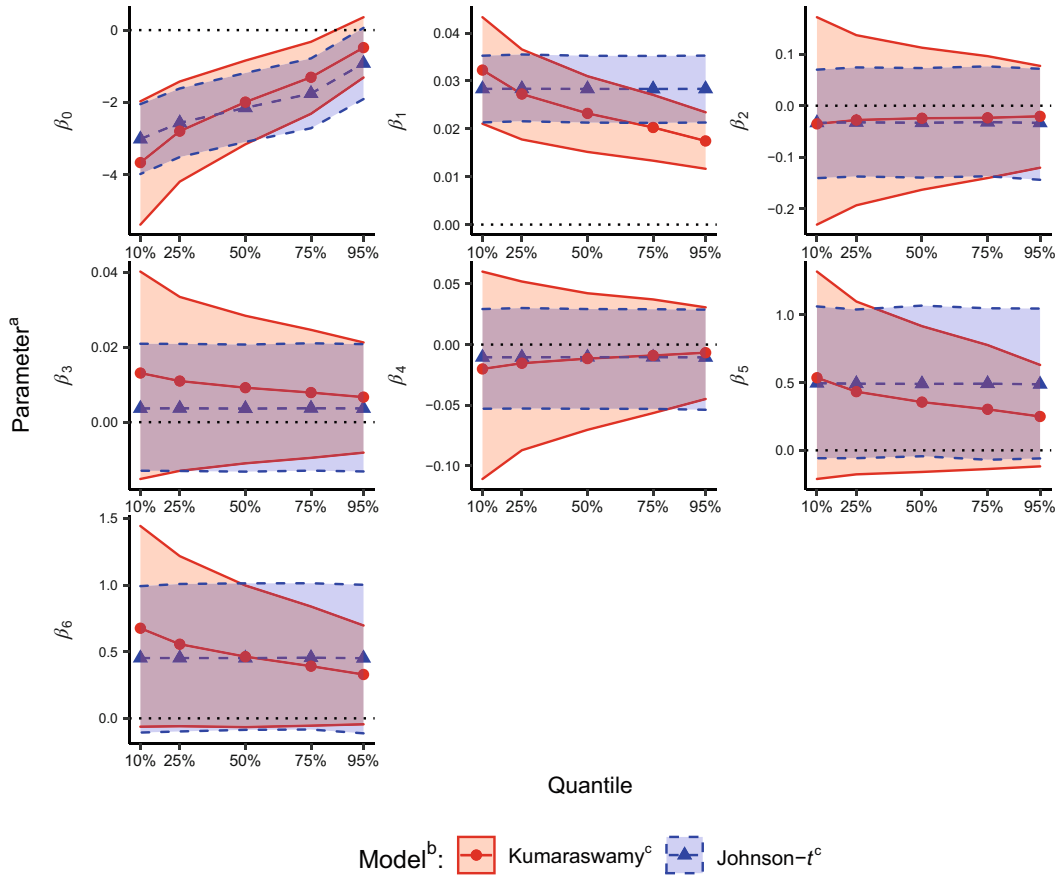

FIGURE 2 Cushion plant dataset: quantile regression coefficients' PEs and 95% HPD intervals computed from the Kumaraswamy and Johnson-t models by parameter, model, and quantile. HPD, highest posterior density. PE, posterior estimate. [a]Predictor: β_1 = Initial dead stem cover (%), β_2 = Initial *Azorella selago* size (cm^2), β_3 = Initial *Agrostis* cover (%), β_4 = Initial cover of other (%), β_5 = Mid versus high, β_6 = West versus east. [b]The red dots and solid lines/shaded bands represent the quantile regression coefficients' PEs and 95% HPD intervals under the Kumaraswamy model, whereas the blue triangles and dash lines/shaded bands represent those under the Johnson-t model; the black dotted line denotes the reference line at zero. [c]The logit link function was used to model the final *A. selago* dead stem cover's quantiles as a function of covariates.

- The effect of the covariates on the quantiles generally differs considerably between the robust and nonrobust models. Under the robust model, the effect of the covariates on the range of quantiles is similar. In contrast, the covariate effects under the nonrobust model differ considerably among the range of quantiles.

Secondly, we compare the fits of the mean and quantile models; we restrict the comparison to the robust models (i.e., the rectangular beta and Johnson-t models). Figure S4 presents the mean and quantile regression coefficients' PEs and 95% HPD intervals calculated from the rectangular beta and Johnson-t models. From Figure S4, we observe that the effect of some covariates on the mean differs somewhat from that of the quantiles.

Thirdly, we compare the fits of the nonrobust and robust mean and median models (i.e., the beta and rectangular beta models for the mean and the Kumaraswamy and Johnson-t models

for the median). The PEs and 95% HPD intervals for the regression coefficients of the mean and median computed from the beta, rectangular beta, Kumaraswamy, and Johnson-*t* models are presented in Figure S5. In addition, the complete set of mean and median model parameters, including the model comparison statistic, are presented in Table S2. From Figure S5 and Table S2, we observe that the robust mean and median models are favored over the nonrobust models and that the PEs and 95% HPD intervals under the nonrobust and robust models differ somewhat. The estimates and confidence intervals of the two central tendency measures' regression coefficients (i.e., the mean and the median) are somewhat different, probably due to skewness in the data.

4.3 | Model adequacy

We determine the influence of observations on model fits using leave-one-out cross-validation to assess model adequacy (Wang & Luo, 2016): in particular, we evaluate the effect of a data point when absent and present in the dataset using the Kullback–Leibler (K-L) divergence (see Section B.1 of Appendix B). We also assess the model residuals and empirical predictive coverage as the goodness of fit measures based on the posterior predictive distribution (see Sections B.2 and B.3 of Appendix B).

Figure S6 in Section A of Data S1 presents the K–L divergence estimates calculated from the beta and rectangular beta models, whereas Figure 3 presents those from the Kumaraswamy and Johnson-*t* models. Many observations considerably affect the estimates of the regression coefficients of mean and quantiles under the nonrobust models (influential outliers), whereas fewer observations significantly affect these estimates under the robust models.

Figures S7 and S8 show the residual diagnostics calculated from the Kumaraswamy and Johnson-*t* models. Under the Kumaraswamy model, the scaled residuals do not vary uniformly between 0 and 1 for all quantiles, implying that the model does not fit the data well. Furthermore, the quantile regression fits of the scaled residuals against the corresponding fitted values deviate considerably from the expected quantiles. In contrast, the residual diagnostics suggest that the Johnson-*t* model fits the data well.

Figure S9 presents the candidate models' empirical predictive coverage and the average length of predictive intervals. The coverage of the predictive intervals from the robust mean and quantile models is close to the nominal value, whereas the predictive coverage under the nonrobust models is too high. The robust mean and quantile models yield narrower predictive intervals than the nonrobust models.

Per the current statistical software implementing the Kumaraswamy model, `R-INLA` (Lindgren & Rue, 2015), we checked the model fits using predictive integral transforms (PITs). The PITs presented in Figure S10 suggest that the Kumaraswamy model fits the data poorly for all investigated quantiles (potentially due to outliers). Therefore, the Kumaraswamy model is not adequate for our dataset.

4.4 | Ecological findings

Since the Johnson-*t* model clearly performs better than the Kumaraswamy model, as shown in the previous section, we quantify the constraints to vitality in cushion plants according to the Johnson-*t* model. We exponentiate the regression coefficients and interpret them as odds ratios, analogous to logistic regression modeling of binomial data. Table 2 summarizes the mean

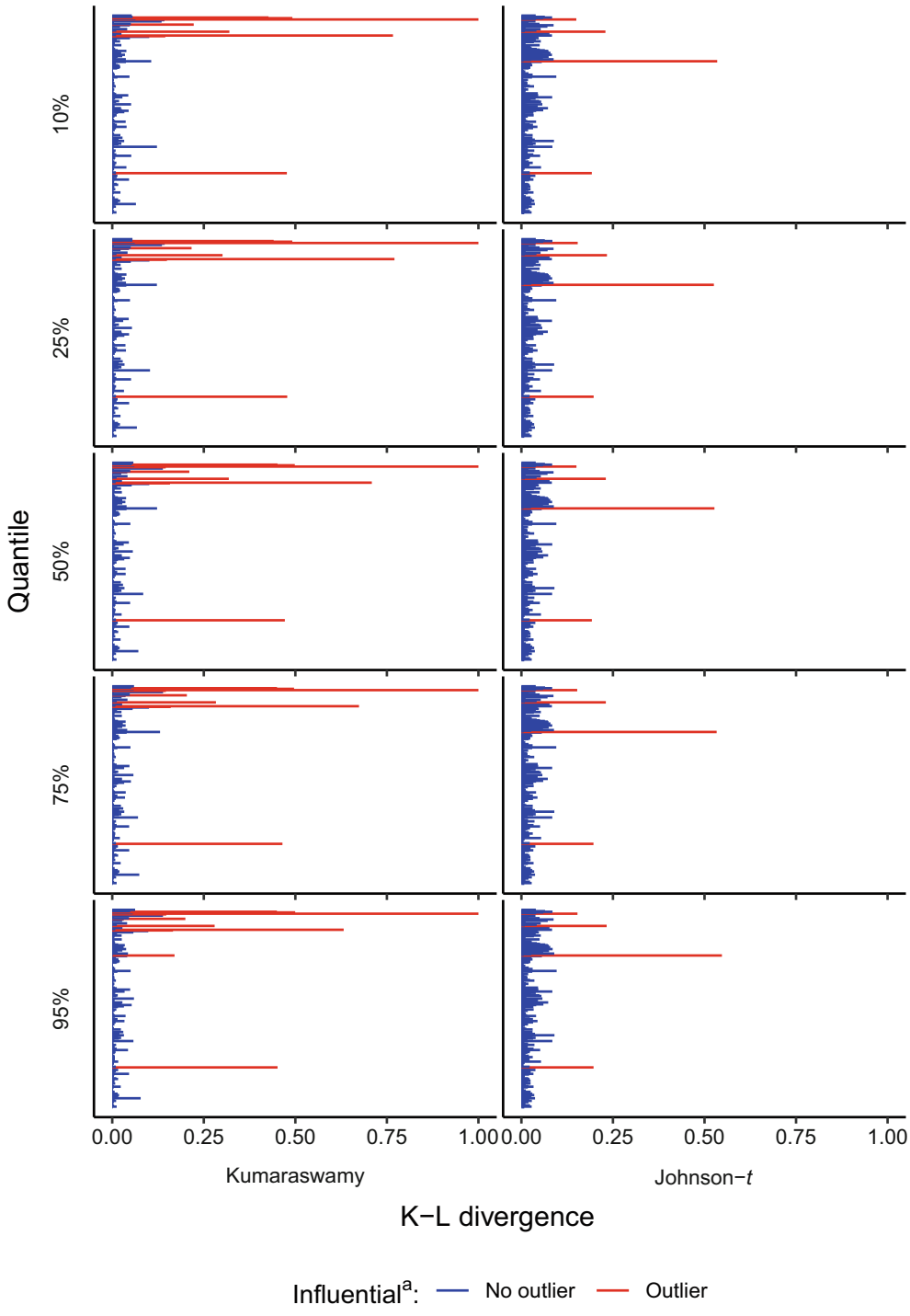

FIGURE 3 Cushion plant dataset: Kullback–Leibler divergence estimates computed from the Kumaraswamy and Johnson-*t* regression models by quantile, observation, and model. [a] The blue lines represent data points that are not considered outliers, whereas the red lines represent influential observations that significantly affect the quantile regression coefficients' estimates (i.e., outliers).

TABLE 2 Cushion plant dataset: mean and quantile regression models' odds ratio estimates computed from the rectangular beta and Johnson-*t* regression model.

		Quantile[b]				
Parameter / Predictor[a]	**Mean[c]**	**10%**	**25%**	**50%**	**75%**	**95%**
β_3 / Initial *Agrostis* cover (%)	1.021	1.018	1.019	1.018	1.019	1.018
β_4 / Initial cover of other (%)	0.897	0.949	0.949	0.949	0.949	0.949
β_5 / Mid versus high	1.502	1.642	1.636	1.634	1.636	1.629
β_6 / West versus east	1.489	1.572	1.573	1.572	1.577	1.571

[a]Predictor: The mean' odds ratios are derived from the rectangular beta distribution.
[b]The odds ratios are calculated according to a 5% increase in the initial cover of vascular and nonvascular plant species growing on the cushion plants (i.e., β_3 and β_4).
[c]The quantiles' odds ratios are derived from the Johnson-*t* distribution.

and quantile models' odds ratio estimates, that is, $\exp(\hat{\beta}_3), \ldots, \exp(\hat{\beta}_6)$, calculated from the rectangular beta and Johnson-*t* models. The estimates of the odds ratios suggest the following:

1. A 5% increase in the initial *Agrostis magellanica* cover results in a 2% increase in the odds of all quantiles of final dead stem cover (see β_3). Similarly, a 5% increase in the initial cover of other vascular plants and mosses results in an approximately 5% decrease in the odds of all quantiles of final dead stem cover (see β_4). Therefore, despite evidence for *A. selago* altering the population structure, biomass, reproductive output, cover, and abundance of *A. magellanica* and other vascular and nonvascular plants compared to surrounding areas where *A. selago* is absent (le Roux, Shaw, & Chown, 2013; Raath-Krüger et al., 2021), the extreme quantiles of vitality in *A. selago* are not constrained by the cover of vascular plants and mosses.
2. The odds of all quantiles of final dead stem cover are about 64% higher for mid-altitude than high-altitude sites (see β_5). Similarly, the odds of all quantiles of final dead stem cover are about 57% higher for the western aspect of Marion Island than the eastern aspect (see β_6). Therefore, the limit to *A. selago* vitality is constrained by altitude and aspect: the odds ratios reveal that, across all quantiles, dead stem cover is higher at the mid-altitude than at high-altitude sites and on the western aspect of the island than the eastern aspect. However, the difference in the extreme quantiles of dead stem cover between mid versus high-altitude sites may depend on *A. magellanica* cover, and therefore, additional covariates may be necessary for the model to account for such interactions.
3. The effect of vascular and nonvascular plant species, altitude, and aspect on *A. selago* does not vary across plants of different vitality. Therefore, it suggests that neither upper nor lower *A. selago* vitality is constrained by environmental conditions (altitude, aspect, and plant cover) differently from the median vitality, a robust central tendency measure.

5 | SIMULATION STUDIES

5.1 | Model performance

We assessed the performance of the Johnson-*t* regression model in Section 3 in a simulation study under two distributional assumptions, namely data generated from the Johnson-*t* distribution

with heavy and light tails, each under three quantile levels, namely the first decile, the median, and the 0.95th quantile (i.e., $q \in \{0.1, 0.5, 0.95\}$; hence a total of six parameter scenarios). The parameter values were chosen to partly reflect the ecology study's typical data in Section 2.

For all scenarios, the sample size and the values of the fixed effects, covariance matrix, and dispersion parameter were chosen as $I = 10$, $J = 5$, $\boldsymbol{x}_{ij} = \boldsymbol{z}_{ij} = \left(1, t_{ij}\right)'$, $\beta_0 = -1.65$, $\beta_1 = 0.05$, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, $\sigma_1^2 = 0.75$, $\sigma_2^2 = 0.7$, $\sigma_{12} = 0.05$, $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$, and $\rho = 1.5$. Note that the model is specified as a random intercept-slope mixed-effects model instead of a random intercepts model as in our application. The degrees of freedom were chosen as (i) $\nu = 5$ (i.e., the heavy-tailed scenario) and (ii) $\nu = 30$ (i.e., the light-tailed scenario). Hence, the following six scenarios were considered in total: (i) $q = 0.1$ and $\nu = 5$, (ii) $q = 0.1$ and $\nu = 30$, (iii) $q = 0.5$ and $\nu = 5$, (iv) $q = 0.5$ and $\nu = 30$, (v) $q = 0.95$ and $\nu = 5$, (vi) $q = 0.95$ and $\nu = 30$. We simulated the t_{ij} from a unit interval uniform distribution, namely Uniform $(0, 1)$. We fitted the Johnson-t model to 500 simulated datasets for each of the six scenarios.

The bias of the PEs was calculated (see details in Appendix C), as was the coverage probability of the associated 95% HPD intervals. We used the `autorun.jags` function of the `runjags` package (Denwood, 2016) to guarantee convergence of the posterior samples for each fitted dataset. The corresponding results are presented in Table 3. From Table 3, we observe the following:

- Under each parameter scenario, the bias of the estimates of the fixed effects (i.e., β_0, β_1) and of the dispersion parameter (i.e., ρ) is small. In contrast, the bias of the estimates of the variance components (i.e., σ_1^2, σ_2^2, σ_{12}) is large.
- The estimates of the degrees of freedom (i.e., ν) are relatively unbiased for low degrees of freedom and considerably biased for high degrees of freedom.
- Under each parameter scenario, the coverage probability of the HPD interval of the fixed effects (i.e., β_0, β_1) is close to or slightly higher than the nominal value. In contrast, the HPD intervals of the dispersion parameter (i.e., ρ) and variance components (i.e., σ_1^2, σ_2^2, σ_{12}) are too conservative.
- When the data are heavy-tailed, the coverage probability of the HPD interval of the degrees of freedom (i.e., ν) is slightly higher than the nominal value. In contrast, when the data are light-tailed, the degrees of freedom's HPD interval coverage probability is somewhat lower than the nominal value.

In summary, the estimates and HPD interval coverage of the fixed effects, which are the main parameters of interest, are acceptable under all parameter scenarios. However, the HPD intervals of the variance components are generally too conservative, similar to the findings of Burger et al. (2020) and Burger and Lesaffre (2021), in particular, under small variance components. The HPD interval coverage of the degrees of freedom is slightly lenient when the data are light-tailed; in contrast, it is acceptable when the data are heavy-tailed.

5.2 | Data contamination

We performed a simulation study to investigate the robustness of the Kumaraswamy and Johnson-t regression models to outliers. Datasets were simulated from the two models where we chose the model parameter values for $I, J, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \beta_0, \beta_1, \sigma_1^2, \sigma_2^2$, and σ_{12} as per Section 5.1. We chose

TABLE 3 Simulation study: performance of the Johnson-t regression model.[a]

			Quantile					
			10%		50%		95%	
ν	Parameter	Value	Bias	Coverage[b]	Bias	Coverage[b]	Bias	Coverage[b]
5	β_0	−1.65	−0.0091	95.0	0.0039	96.2	0.0718	96.0
	β_1	0.05	0.0060	96.0	−0.0530	97.2	0.0008	97.8
	ρ	1.5	0.0524	98.0	0.0627	98.8	0.0541	99.2
	ν	5	1.4533	97.4	1.3483	98.0	1.2336	97.2
	σ_1^2	0.75	0.1813	97.2	0.1546	98.2	0.1523	97.0
	σ_2^2	0.7	0.3158	99.8	0.3473	100.0	0.4039	99.6
	σ_{12}	0.05	−0.0818	100.0	−0.0712	100.0	−0.0761	100.0
30	β_0	−1.65	−0.0085	97.2	0.0005	95.0	0.0674	96.6
	β_1	0.05	0.0090	97.2	−0.0192	96.6	−0.0283	97.0
	ρ	1.5	0.1929	98.4	0.2144	97.0	0.2251	98.2
	ν	30	−21.6977	93.0	−21.6871	92.4	−21.6852	92.2
	σ_1^2	0.75	0.1702	95.6	0.1492	97.0	0.1641	98.6
	σ_2^2	0.7	0.2412	100.0	0.2066	100.0	0.2422	99.8
	σ_{12}	0.05	-0.0525	100.0	−0.0615	100.0	−0.0584	100.0

[a]The logit link function was used to model the response variable's quantiles as a function of covariates.
[b]Coverage of 95% highest posterior density intervals (%).

$\nu = 30$, and $\rho = 3$ and $\rho = 1.5$ for data simulated from the Kumaraswamy and Johnson-t models, respectively. The selection of the degrees of freedom, precision, and dispersion parameter values yields comparable datasets, ensuring a sensible comparison between the two candidate models.

The datasets were randomly contaminated by replacing y_{ij} with data simulated from the Uniform (0.975, 1) distribution at a rate of 5%, thereby contaminating the data with outliers close to the upper bound of the parameter space (i.e., 1). Hence, our contamination scheme is based on a mixture of the Kumaraswamy/Johnson-t distribution and the uniform distribution. The candidate models were fitted to both the uncontaminated and contaminated versions of the simulated datasets. Under each scenario, we considered two quantile levels: the median and the 0.95th quantile (i.e., $q \in \{0.5, 0.95\}$; hence a total of four parameter scenarios for each model). We fitted the models to 500 simulated datasets for each scenario.

The bias and root mean square error (RMSE) of the PEs were calculated (see details in Appendix C), as were the average length and empirical coverage probability of the associated 95% HPD intervals. The corresponding results are presented in Table 4 (fixed effects only). From Table 4, we observe the following:

- Both models perform well under no contamination, as judged by the accuracy characteristic (i.e., bias) and HPD interval coverage.
- Under "contamination" relative to "no contamination":
  - The Kumaraswamy model's median coefficient estimates are more biased and less precise, most notably the fixed intercept term; in contrast, the bias under the Johnson-t model is small.

TABLE 4 Simulation study: robustness of the Kumaraswamy and Johnson-t regression models.

			Kumaraswamy[a]					Johnson-t[a]				
Quantile	**Rate[b]**	**Parameter**	**Value**	**Bias**	**RMSE**	**Coverage[c]**	**Length[d]**	**Value**	**Bias**	**RMSE**	**Coverage[c]**	**Length[d]**
50%	0%	β_0	−1.65	0.0218	0.3667	97.2	1.6803	−1.65	0.0005	0.3609	95.0	1.5662
		β_1	0.05	0.0066	0.5338	98.0	2.4134	0.05	−0.0192	0.4726	96.6	2.1089
		ρ	3	0.0493	0.4035	96.4	1.5602	1.5	0.2144	0.3416	97.0	1.3529
		ν	NA	NA	NA	NA	NA	30	−21.6871	21.8615	92.4	69.8803
		σ_1^2	0.75	0.1505	0.6229	98.4	2.8504	0.75	0.1492	0.5745	97.0	2.6462
		σ_2^2	0.7	0.3101	0.9635	99.6	4.6387	0.7	0.2066	0.8026	100.0	3.9073
		σ_{12}	0.05	−0.0628	0.1820	100.0	2.3097	0.05	−0.0615	0.1876	100.0	2.0335
	5%	β_0	−1.65	0.3376	0.5792	95.6	2.2589	−1.65	0.0132	0.3749	95.6	1.6235
		β_1	0.05	0.0387	0.6927	100.0	3.8310	0.05	0.0021	0.5192	96.2	2.2973
		ρ	3	−1.7550	1.8600	8.0	0.7490	1.5	0.5991	0.7732	93.2	2.3549
		ν	NA	NA	NA	NA	NA	30	−27.8206	27.8844	6.0	7.2147
		σ_1^2	0.75	−0.1247	1.4820	97.4	2.9727	0.75	0.1411	0.5813	97.4	2.7404
		σ_2^2	0.7	0.6553	5.1296	99.2	7.2064	0.7	0.2793	1.0227	100.0	4.4266
		σ_{12}	0.05	−0.3355	2.3501	99.4	3.2042	0.05	−0.0669	0.1953	100.0	2.1908

(Continues)

TABLE 4 Continued

			Kumaraswamy[a]					Johnson-t[a]				
Quantile	**Rate[b]**	**Parameter**	**Value**	**Bias**	**RMSE**	**Coverage[c]**	**Length[d]**	**Value**	**Bias**	**RMSE**	**Coverage[c]**	**Length[d]**
95%	0%	β_0	−1.65	0.0040	0.3441	97.0	1.5913	−1.65	0.0674	0.3830	96.6	1.7489
		β_1	0.05	0.0198	0.4291	95.2	1.9443	0.05	−0.0283	0.4761	97.0	2.1281
		ρ	3	0.0618	0.5713	95.6	2.2042	1.5	0.2251	0.3682	98.2	1.3766
		ν	NA	NA	NA	NA	NA	30	−21.6852	21.8669	92.2	69.7852
		σ_1^2	0.75	0.2027	0.5649	97.0	2.7025	0.75	0.1641	0.5701	98.6	2.6989
		σ_2^2	0.7	0.2848	0.9156	98.8	3.7817	0.7	0.2422	0.8183	99.8	4.0036
		σ_{12}	0.05	−0.0513	0.2440	100.0	2.0793	0.05	−0.0584	0.1995	100.0	2.1002
	5%	β_0	−1.65	2.2985	2.5372	10.0	2.0452	−1.65	0.7663	0.9692	85.4	2.9303
		β_1	0.05	0.0080	0.3795	99.4	2.3732	0.05	−0.0322	0.5221	96.6	2.2980
		ρ	3	−2.5188	2.6458	8.2	0.4832	1.5	0.6845	0.8712	90.8	2.4950
		ν	NA	NA	NA	NA	NA	30	−27.8940	27.9721	7.2	7.6088
		σ_1^2	0.75	−0.4378	0.8387	58.2	1.2959	0.75	0.1747	0.5910	98.8	2.8162
		σ_2^2	0.7	−0.0394	2.1986	98.0	3.1144	0.7	0.2716	0.8575	100.0	4.4121
		σ_{12}	0.05	−0.1519	0.9710	99.2	1.2521	0.05	−0.0659	0.1932	100.0	2.2635

Abbreviations: HPD, Highest posterior density; NA, not applicable.

[a]The logit link function was used to model the response variable's quantiles as a function of covariates.

[b]Contamination rate.

[c]95% HPD interval coverage (%).

[d]95% HPD interval average length.

- The 0.95th quantile's coefficient estimates are considerably more biased and less precise under both models. However, the lack of precision of the estimates of the intercept term is more prominent for the Kumaraswamy model.
- The HPD interval of the median slope term is extremely conservative under the Kumaraswamy model, whereas the coverage of the HPD interval of the median slope term from the Johnson-*t* model is acceptable.
- The 0.95th quantile's HPD interval of the intercept term is extremely lenient under the Kumaraswamy model; in contrast, the HPD interval's lack of coverage is considerably less problematic under the Johnson-*t* model.
- The increase in the HPD intervals' average length under the Johnson-*t* model is generally considerably less than that of the Kumaraswamy model.

In summary, the data contamination simulation study suggests that the Johnson-*t* model is considerably more robust to outliers than the Kumaraswamy model.

6 | DISCUSSION

The currently available approach for modeling the quantiles of hierarchically structured continuous proportion data is the Kumaraswamy model of Bayes et al. (2017). However, our application dataset contains significant outliers, and the Kumaraswamy model is not adequate for modeling heavy-tailed data (i.e., containing outliers). We, therefore, considered a robust model to accommodate outliers by extending the fixed effects Johnson-*t* model of Lemonte and Moreno-Arenas (2020). According to the mDIC statistic, the robust models (i.e., the rectangular beta and Johnson-*t* models) fit the cushion plant dataset better than the nonrobust models (i.e., the beta and Kumaraswamy models). Furthermore, the quantile model fits differ considerably between the nonrobust and robust models. Under the robust models, the covariate effects on the mean and median differ somewhat; therefore, one should carefully consider which measure of central tendency to report (i.e., mean vs. median) for bounded data. Based on the model adequacy checks, it is clear that the outliers in the data have a substantial effect on the quantile fits, and thus, the ecological research questions were addressed using the Johnson-*t* model instead of the Kumaraswamy model.

We used a Bayesian implementation of the models considered for this manuscript. We chose the specification of the MGH-*t* prior for the variance components (Huang & Wand, 2013) over the conventional Wishart prior since the latter may yield too lenient confidence intervals than the former. Overly lenient confidence intervals for the variance components can be much more problematic for inferences about fixed effects than extremely conservative confidence intervals for the variance components (Burger et al., 2020).

The data contamination simulation study suggests that the Johnson-*t* model for modeling the quantiles of continuous proportions is remarkably robust to outliers, whereas the Kumaraswamy model is susceptible to outliers, especially when modeling the extreme quantiles.

The HPD intervals of the variance components are generally very conservative; however, the coverage for the fixed effects is satisfactory. The degrees of freedom estimates are biased when there are no outliers in the data. However, accurate estimation of the *t*-distribution's degrees of freedom is known to be challenging: see Fonseca, Ferreira, and Migon (2008) for a detailed explanation; furthermore, in the current context, precise estimation of the degrees of freedom is not required. Instead, we need to determine whether the degrees of freedom are (i) small, which leads

to a model robust to outliers, or (ii) large, suggesting that outliers are not an issue. Overall, the simulation study suggests that the proposed robust model has good accuracy and confidence interval coverage properties.

The Jeffreys prior of Juárez and Steel (2010) for the degrees of freedom can be used as an alternative to the hierarchical prior as a sensitivity analysis. However, the trigamma function is not available in `JAGS`; the `Stan` software (Carpenter et al., 2017) can alternatively be used to specify the Jeffreys prior as it contains the trigamma function.

It should be noted that, in some cases, excluding outliers because they deviate considerably from the other observations may misrepresent vital ecological processes, ultimately leading to misleading conclusions. From a statistical perspective, it seems preferable to carry out an analysis that is robust to outliers rather than an analysis that is preceded by outlier removal.

The robust regression models introduced can be extended to model the precision and dispersion parameters as a function of covariates.

Should inferences about the quantiles of the continuous proportions on the population level be of interest (i.e., as opposed to conditional on the random effects), additional computationally expensive steps to integrate (marginalize) over the distribution of the random effects are needed for our proposed model.

In conclusion, our study demonstrated that the proposed Johnson-*t* model is an appropriate robust alternative to the current approach, the Kumaraswamy model, for modeling the quantiles of correlated continuous proportions when outliers are present in the data.

ACKNOWLEDGMENTS

The authors wish to acknowledge Melodie McGeoch for providing the photographs from which the 2003 data were extracted for Raath-Krüger et al. (2022). The data were obtained as part of research supported by the National Research Foundation (NRF) via the South African National Antarctic Programme (Grant numbers 93077 and 110726). We also thank Janet van Niekerk, Mohammad Arashi, Andriëtte Bekker, Robert Schall, and Ashenafi Yirga for discussions that improved the quality of the manuscript. This work is based on the research supported by the NRF of South Africa (Grant number 132383 and Postdoctoral Grant PDG190329424983). Opinions expressed, and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Divan A. Burger https://orcid.org/0000-0001-8096-6371
Emmanuel Lesaffre https://orcid.org/0000-0002-3747-6905
Peter C. le Roux https://orcid.org/0000-0002-7941-7444
Morgan J. Raath-Krüger https://orcid.org/0000-0002-3326-1165

REFERENCES

Bayes, C. L., Bazán, J. L., & de Castro, M. (2017). A quantile parametric mixed regression model for bounded response variables. *Statistics and Its Interface*, *10*(3), 483–493.

Bayes, C. L., Bazán, J. L., & García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis*, *7*(4), 841–866.

Begashaw, G. B., & Yohannes, Y. B. (2020). Review of outlier detection and identifying using robust regression model. *International Journal of Systems Science and Applied Mathematics*, *5*(1), 4.

Benhadi-Marín, J. (2018). A conceptual framework to deal with outliers in ecology. *Biodiversity and Conservation*, *27*(12), 3295–3300.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., … Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Burger, D. A., & Lesaffre, E. (2021). Nonlinear mixed-effects modeling of longitudinal count data: Bayesian inference about median counts based on the marginal zero-inflated discrete Weibull distribution. *Statistics in Medicine*, *40*(23), 5078–5095.

Burger, D. A., Schall, R., Ferreira, J. T., & Chen, D.-G. (2020). A robust Bayesian mixed effects approach for zero inflated and highly skewed longitudinal count data emanating from the zero inflated discrete Weibull distribution. *Statistics in Medicine*, *39*(9), 1275–1291. https://doi.org/10.1002/sim.8475

Cade, B. S., & Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, *1*(8), 412–420.

Cancho, V. G., Bazán, J. L., & Dey, D. K. (2020). A new class of regression model for a bounded response with application in the study of the incidence rate of colorectal cancer. *Statistical Methods in Medical Research*, *29*(7), 2015–2033.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.

Denwood, M. J. (2016). Runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(9), 1–25.

di Brisco, A. M., & Migliorati, S. (2020). A new mixed-effects mixture model for constrained longitudinal data. *Statistics in Medicine*, *39*(2), 129–145.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815.

Flores, S. E., Prates, M. O., Bazán, J. L., & Bolfarine, H. B. (2021). Spatial regression models for bounded response variables with evaluation of the degree of dependence. *Statistics and Its Interface*, *14*(2), 95–107.

Fonseca, T. C. O., Ferreira, M. A. R., & Migon, H. S. (2008). Objective Bayesian analysis for the student-*t* regression model. *Biometrika*, *95*(2), 325–333.

Galarza, C. E., Zhang, P., & Lachos, V. H. (2020). Logistic quantile regression for bounded outcomes using a family of heavy-tailed distributions. *Sankhya B*, *83*, 325–349.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409. https://doi.org/10.1080/01621459.1990.10476213

Gelman, A., & Hill, J. (Eds.). (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Goddard, K. A., Craig, K. J., Schoombie, J., & le Roux, P. C. (2022). Investigation of ecologically relevant wind patterns on Marion Island using computational fluid dynamics and measured data. *Ecological Modelling*, *464*, 109827.

Hartig, F. (2021a). *DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models*. Retrieved from. https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html

Hartig, F. (2021b). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. R Package Version 0.4.4.

Huang, A., & Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, *8*(2), 439–452. https://doi.org/10.1214/13-BA815

Huntley, B. J. (1972). Notes on the ecology of Azorella selago hook. f. *Journal of South African Botany*, *38*, 103–113.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, *36*, 149–176.

Juárez, M. A., & Steel, M. F. J. (2010). Model-based clustering of non-Gaussian panel data based on skew-t distributions. *Journal of Business & Economic Statistics*, *28*(1), 52–66.

Kellner, K. (2021). jagsUI: A wrapper around rjags to streamline JAGS analyses. *R Package Version*, *1*(5), 2.

Koenker, R., & Bassett, G., Jr. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.

Koenker, R., Portnoy, S., Ng, P. T., Melly, B., Zeileis, A., Grosjean, P., … Ripley, B. D. (2021). *Quantreg: Quantile regression*. R package version 5.86.

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, *70*(4), 407.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association*, *84*(408), 881–896.

le Roux, P. C. (2008). *Climate and climate change*. In S. L. Chown & P. W. Froneman (Eds.), *The Prince Edward islands: Land-Sea interactions in a changing ecosystem* (pp. 39–64). Stellenbosch, South Africa: African Sun-Media.

le Roux, P. C., McGeoch, M. A., Nyakatya, M. J., & Chown, S. L. (2005). Effects of a short-term climate change experiment on a sub-Antarctic keystone plant species. *Global Change Biology*, *11*(10), 1628–1639.

le Roux, P. C., Shaw, J. D., & Chown, S. L. (2013). Ontogenetic shifts in plant interactions vary with environmental severity and affect population structure. *New Phytologist*, *200*(1), 241–250.

Lemonte, A. J., & Moreno-Arenas, G. (2020). On a heavy-tailed parametric quantile regression model for limited range response variables. *Computational Statistics*, *35*(1), 379–398.

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150–156.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*(1), 1–25.

Mazucheli, J., Leiva, V., Alves, B., & Menezes, A. F. B. (2021). A new quantile regression for modeling bounded data under a unit Birnbaum–Saunders distribution with applications in medicine and politics. *Symmetry*, *13*(4), 682.

Mazucheli, J., Menezes, A. F. B., Fernandes, L. B., de Oliveira, R. P., & Ghitany, M. E. (2020). The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modeling of quantiles conditional on covariates. *Journal of Applied Statistics*, *47*(6), 954–974.

McGeoch, M. A., le Roux, P. C., Hugo, E. A., & Nyakatya, M. J. (2008). *Spatial variation in the terrestrial biotic system*. In S. L. Chown & P. W. Froneman (Eds.), *The Prince Edward islands: Land-Sea interactions in a changing ecosystem* (pp. 245–276). Stellenbosch, South Africa: African SunMedia.

Migliorati, S., di Brisco, A. M., & Ongaro, A. (2018). A new regression model for bounded responses. *Bayesian Analysis*, *13*(3), 845–872.

Min, I., & Kim, I. (2004). A Monte Carlo comparison of parametric and nonparametric quantile regressions. *Applied Economics Letters*, *11*(2), 71–74.

Mitnik, P. A., & Baek, S. (2013). The Kumaraswamy distribution: Median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Statistical Papers*, *54*(1), 177–192.

Nyakatya, M. J., & McGeoch, M. A. (2008). Temperature variation across Marion Island associated with a keystone plant species (Azorella selago Hook.(Apiaceae)). *Polar Biology*, *31*(2), 139–151.

Owen, W. R. (1995). Growth and reproduction in an alpine cushion plant: Astragalus kentrophyta var. implexus. *The Great Basin Naturalist*, *55*, 117–123.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Paper presented at the meeting of the Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria, 1–10.

Quintero, A., & Lesaffre, E. (2018). Comparing hierarchical models via the marginalized deviance information criterion. *Statistics in Medicine*, *37*(16), 2440–2454. https://doi.org/10.1002/sim.7649

Raath-Krüger, M. J., Schöb, C., McGeoch, M. A., Burger, D. A., Strydom, T., & le Roux, P. C. (2022). Long-term spatially-replicated data show no cost to a benefactor species in a facilitative plant-plant interaction. *bioRxiv*. Retrieved from. https://www.biorxiv.org/content/10.1101/2022.10.17.512641v1

Raath-Krüger, M. J., Schöb, C., McGeoch, M. A., & le Roux, P. C. (2021). Interspecific facilitation mediates the outcome of intraspecific interactions across an elevational gradient. *Ecology*, *102*(1), e03200.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*(1), 1–28.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, *64*(4), 583–639. https://doi.org/10.1111/1467–9868.00353

Tomazella, V. L. D., Jesus, S. R., Gazon, A. B., Louzada, F., Nadarajah, S., Nascimento, D. C., … Ramos, P. L. (2021). Bayesian reference analysis for the generalized normal linear regression model. *Symmetry*, *13*(5), 856.

Wang, J., & Luo, S. (2016). Augmented Beta rectangular regression models: A Bayesian perspective. *Biometrical Journal*, *58*(1), 206–221.

Wei, Y., Kehm, R. D., Goldberg, M., & Terry, M. B. (2019). Applications for quantile regression in epidemiology. *Current Epidemiology Reports*, *6*(2), 191–199.

Yirga, A. A., Melesse, S. F., Mwambi, H. G., & Ayele, D. G. (2021). Additive quantile mixed effects modelling with application to longitudinal CD4 count data. *Scientific Reports*, *11*(1), 1–12.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Burger, D. A., van der Merwe, S., Lesaffre, E., le Roux, P. C., & Raath-Krüger, M. J. (2023). A robust mixed-effects parametric quantile regression model for continuous proportions: Quantifying the constraints to vitality in cushion plants. *Statistica Neerlandica*, *77*(4), 444–470. https://doi.org/10.1111/stan.12293

APPENDIX A. MODEL DISCRIMINATION

Let $\boldsymbol{\Psi} = \left(\boldsymbol{\beta}', \rho\right)'$ under the conventional beta and Kumaraswamy regression models, $\boldsymbol{\Psi} = \left(\boldsymbol{\beta}', \rho, \phi\right)'$ under the rectangular beta model, and $\boldsymbol{\Psi} = \left(\boldsymbol{\beta}', \rho, \nu\right)'$ under the Johnson-t model. Furthermore, let $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\Sigma}}$, respectively, represent the posterior estimate of $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$.

The mDIC requires integrating over the likelihood function's random effects. In order to do so, Quintero and Lesaffre (2018) proposed generating replicate samples of the random effects that need to be integrated out. Two sources of replicate samples of the random effects are drawn as follows: let $\boldsymbol{\Psi}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$, respectively, represent the kth posterior sample drawn from $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$, $\boldsymbol{u}_{\text{rep}_i}^{(k,l)}$ the l^{th} replicate sample from $p\left(\boldsymbol{u}_i \middle| \boldsymbol{\Sigma}^{(k)}\right)$, and $\boldsymbol{u}_{\text{rep}_i}^{(m)}$ the mth replicate sample from $p\left(\boldsymbol{u}_i \middle| \hat{\boldsymbol{\Sigma}}\right)$ ($k = 1, \ldots, K$, $l = 1, \ldots, L$, and $m = 1, \ldots, M$). Accordingly, the mDIC statistic is calculated under regression model R as follows:

$$\begin{aligned}\text{mDIC}\left(R\right) = &-\frac{4}{K}\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\log\left\{\frac{1}{L}\sum_{l=1}^{L}f\left(y_{ij}\middle|\boldsymbol{\Psi}^{(k)}, \boldsymbol{u}_{\text{rep}_i}^{(k,l)}\right)\right\}\\ &-2\sum_{i=1}^{I}\sum_{j=1}^{J_i}\log\left\{\frac{1}{M}\sum_{m=1}^{M}f\left(y_{ij}\middle|\hat{\boldsymbol{\Psi}}, \boldsymbol{u}_{\text{rep}_i}^{(m)}\right)\right\}.\end{aligned}$$

We refer the reader to Quintero and Lesaffre (2018) to appropriately choose the number of replications M and L. We drew $M = 10{,}000$ and $L = 1000$ replicate samples of the random effects to calculate the mDIC statistic, which is efficient for relatively small datasets containing in the range of 500 observations or less such as the dataset of Raath-Krüger et al. (2022). Note that the model with the smallest mDIC is favored.

APPENDIX B. MODEL ADEQUACY

B.1 Kullback–Leibler divergence

In addition to Appendix A's notation and definitions, let $\boldsymbol{u}_i^{(k)}$ represent the kth posterior sample drawn from $\boldsymbol{u}_i$, $\boldsymbol{\Theta}$ the full set of model parameters, and $\boldsymbol{y}$ the vector containing y_{ij} for all $i = 1, \ldots, I$ and $j = 1, \ldots, J_i$. Furthermore, let $P(\boldsymbol{\Theta}|\boldsymbol{y})$ represent the posterior distribution of $\boldsymbol{\Theta}$ for all $\boldsymbol{y}$ (complete dataset), and $P\left(\boldsymbol{\Theta}\middle|\boldsymbol{y}_{[ij]}\right)$ the posterior distribution of $\boldsymbol{\Theta}$ with observation y_{ij} excluded. The Monte Carlo estimate of the K-L divergence between $P(\boldsymbol{\Theta}|\boldsymbol{y})$ and $P\left(\boldsymbol{\Theta}\middle|\boldsymbol{y}_{[ij]}\right)$ under regression model R is given by:

$$\mathrm{KL}_R\left[P(\boldsymbol{\Theta}|\boldsymbol{y}), P\left(\boldsymbol{\Theta}\middle|\boldsymbol{y}_{[ij]}\right)\right] = \log\left\{\frac{1}{K}\sum_{k=1}^{K}\left[f\left(y_{ij}\middle|\boldsymbol{\Psi}^{(k)}, \boldsymbol{u}_i^{(k)}\right)\right]^{-1}\right\} + \frac{1}{K}\sum_{k=1}^{K}\log\left[f\left(y_{ij}\middle|\boldsymbol{\Psi}^{(k)}, \boldsymbol{u}_i^{(k)}\right)\right].$$

We calculate the K-L divergence estimates for each observation i and j to determine whether y_{ij} under regression model R is influential, that is, identifying data points that significantly affect the parameter estimates. Following the approach of Tomazella et al. (2021), we consider y_{ij} an outlier if $0.5\left(1 + \sqrt{1 - \exp\left(-2KL_R\left[P(\boldsymbol{\Theta}|\boldsymbol{y}), P\left(\boldsymbol{\Theta}\middle|\boldsymbol{y}_{[ij]}\right)\right]\right)}\right) \geq 0.75$.

B.2 Residual analysis

In applying hierarchical mixed-effects regression modeling, the use of standard residual plots as diagnostic tools are limited in reliably identifying model misspecification and, in some cases, may imply that the model fits the data poorly even if the model is correctly specified. Therefore, to circumvent the limitations associated with the inspection of raw residuals for assessing model adequacy, we use the simulation-based strategy that Hartig (2021a) proposes, scaling the residuals between zero and one. In particular, we base our assessment of the residuals on the posterior predictive distribution corresponding to each observation in our dataset, conditional on the associated random effects. For each observation and under each model, we simulate the posterior predictive distribution corresponding to the observation, which allows us to compare an observed value to what we expect it to be in a probabilistic sense. In other words, if the model appropriately fits the data, the predicted values corresponding to the observed values will be close. For the kth set of posterior samples drawn, we draw a random copy $y_{\mathrm{rep}_{ij}}^{(k)}$ of $f\left(y_{ij}\middle|\boldsymbol{\Psi}^{(k)}, \boldsymbol{u}_i^{(k)}\right)$. The generated values $\boldsymbol{y}_{\mathrm{rep}_{ij}} = \left(y_{\mathrm{rep}_{ij}}^{(1)}, \ldots, y_{\mathrm{rep}_{ij}}^{(k)}, \ldots, y_{\mathrm{rep}_{ij}}^{(K)}\right)'$ represent samples from the posterior predictive distribution corresponding to y_{ij}.

We calculate the scaled residuals r_{ij} as the value of the posterior predictive distribution's empirical density function evaluated at the observed value y_{ij}. Hence, the derived residuals are between 0 and 1. A residual equal to (i) 0 indicates that the observed value is below all expectations from the model; (ii) 0.5 indicates that it fits in the middle of what was expected; and (iii) 1 indicates that the value is above all expectations from the model. Should a model fit perfectly, we expect the scaled residuals to vary approximately uniformly between 0 and 1. This work implements the residual analysis similar to that available in the `DHARMa` package (Hartig, 2021b). The motivation and workings of the package (extending the methods of Dunn and Smyth (1996) and Gelman and Hill (2006)) are explained in detail by Hartig (2021a).

The residuals can be checked on a visual basis using (i) a QQ plot of the scaled residuals r_{ij} compared to the theoretical uniform distribution and (ii) a plot of the scaled residuals against the corresponding fitted values, namely, r_{ij} versus $\hat{\kappa}_{ij} = \frac{\exp\left(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}+\boldsymbol{z}'_{ij}\hat{\boldsymbol{u}}_i\right)}{1+\exp\left(\boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}}+\boldsymbol{z}'_{ij}\hat{\boldsymbol{u}}_i\right)}$. We superimposed the quantile regression fits of r_{ij} against $\hat{\kappa}_{ij}$ on the "residual versus prediction" plot to assess whether the r_{ij} are uniformly spread over the $\hat{\kappa}_{ij}$; we used the `R` package `quantreg` (Koenker et al., 2021) to calculate the quantile regression fits.

B.3 Empirical predictive coverage

We calculate posterior predictive intervals using the same principle of comparing observations to their corresponding posterior predictive distributions in Section B.2. We compare the intervals to the observed values by deriving the proportion of the observed values covered by the predictive intervals: this represents an empirical estimate of predictive coverage. We expect a perfectly fitted model to yield empirical predictive coverage close to the nominal coverage. We also calculate the predictive intervals' average length. The model that produces the shortest intervals while maintaining accurate coverage is considered to make better predictions. We studied the interval coverage and lengths for target coverages ranging from 0.01 to 0.99 using increments of 0.02.

APPENDIX C. BIAS AND PRECISION

The bias of a certain estimator E for parameter ϵ is calculated as follows:

$$\text{BIAS} = \frac{\sum_{s=1}^{S}(E_s - \epsilon)}{S},$$

where E_s is the PE calculated in the sth simulation, and the summation is over the S simulations carried out. Similarly, the RMSE is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{s=1}^{S}(E_s - \epsilon)^2}{S}}.$$

An estimator's bias is used to assess its accuracy, while the RMSE serves as a combined measure of accuracy and precision.