









## RESEARCH ARTICLE

**REVISED** **Introducing a drift and diffusion framework for childhood growth research [version 2; peer review: 2 approved]**

Fraser I Lewis <sup>1</sup>, Godfrey Guga <sup>2</sup>, Paschal Mdoe<sup>2</sup>, Esto Mduma <sup>2</sup>, Cloupas Mahopo <sup>3</sup>, Pascal Bessong <sup>3</sup>, Stephanie A Richard<sup>4</sup>, Benjamin J J McCormick <sup>4</sup>

<sup>1</sup>Independent Researcher, Utrecht, The Netherlands

<sup>2</sup>Haydom Lutheran Hospital, Haydom, Tanzania

<sup>3</sup>University of Venda, Thohoyandou, 0950, South Africa

<sup>4</sup>Fogarty International Center, Bethesda, MD, USA

**v2** **First published:** 29 Jun 2020, 4:71  
<https://doi.org/10.12688/gatesopenres.13123.1>  
**Latest published:** 26 Nov 2020, 4:71  
<https://doi.org/10.12688/gatesopenres.13123.2>

**Abstract**

**Background:** Growth trajectories are highly variable between children, making epidemiological analyses challenging both to the identification of malnutrition interventions at the population level and also risk assessment at individual level. We introduce stochastic differential equation (SDE) models into child growth research. SDEs describe flexible dynamic processes comprising: drift - gradual smooth changes - such as physiology or gut microbiome, and diffusion - sudden perturbations, such as illness or infection.







**Methods:** We present a case study applying SDE models to child growth trajectory data from the Haydom, Tanzania and Venda, South Africa sites within the MAL-ED cohort. These data comprise n=460 children aged 0-24 months. A comparison with classical curve fitting (linear mixed models) is also presented.


**Results:** The SDE models offered a wide range of new flexible shapes and parameterizations compared to classical additive models, with performance as good or better than standard approaches. The predictions from the SDE models suggest distinct longitudinal clusters that form distinct 'streams' hidden by the large between-child variability.

**Conclusions:** Using SDE models to predict future growth trajectories revealed new insights in the observed data, where trajectories appear to cluster together in bands, which may have a future risk assessment application. SDEs offer an attractive approach for child growth modelling and potentially offer new insights.

**Open Peer Review**

**Reviewer Status**  

	Invited Reviewers	
	1	2
<b>version 2</b>		
(revision)		
26 Nov 2020	report	report
		
<b>version 1</b>		
29 Jun 2020		
	report	report

1. **Noel Cameron**, Loughborough University, Loughborough, UK
2. **James A. Berkley** , The Childhood Acute Illness and Nutrition (CHAIN) Network, Nairobi, Kenya  
University of Oxford, Oxford, UK  
KEMRI/Wellcome Trust Research Programme, Kilifi, Kenya

Any reports and responses or comments on the article can be found at the end of the article.

## Keywords

child growth, stochastic differential equations, dynamic modelling, MAL-ED

**Corresponding author:** Benjamin J J McCormick ([ben@science.fish](mailto:ben@science.fish))

**Author roles:** **Lewis FI:** Conceptualization, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation; **Guga G:** Data Curation, Writing – Review & Editing; **Mdoe P:** Writing – Review & Editing; **Mduma E:** Writing – Review & Editing; **Mahopo C:** Data Curation, Writing – Review & Editing; **Bessong P:** Writing – Review & Editing; **Richard SA:** Investigation, Writing – Review & Editing; **McCormick BJJ:** Investigation, Methodology, Supervision, Visualization, Writing – Original Draft Preparation

**Competing interests:** Dr. Lewis is an employee of Danone Nutricia Research. All other authors declare no potential conflict of interest.

**Grant information:** This work was supported by the Bill and Melinda Gates Foundation [OPP47075] in collaboration with the Foundation for the NIH and the National Institutes of Health, Fogarty International Center, through support for the Etiology, Risk Factors and Interactions of Enteric Infections and Malnutrition and the Consequences for Child Health and Development Project (MAL-ED). *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2020 Lewis FI *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Lewis FI, Guga G, Mdoe P *et al.* **Introducing a drift and diffusion framework for childhood growth research [version 2; peer review: 2 approved]** Gates Open Research 2020, 4:71 <https://doi.org/10.12688/gatesopenres.13123.2>

**First published:** 29 Jun 2020, 4:71 <https://doi.org/10.12688/gatesopenres.13123.1>

**REVISED Amendments from Version 1**

An error in the coding of the final figure was noted and corrected. Examples of additional factors that could be considered as aspects of drift and diffusion are noted and some additional clarity about what the SDE approach brings.

**Any further responses from the reviewers can be found at the end of the article**

## Introduction

Assessing and analysing growth is a key activity in paediatric epidemiology, building on centuries of research<sup>1</sup>. Anthropometrics are easy to measure with basic equipment and the results are both immediate and meaningful with standardised reference measurements representative of unconstrained growth available from the World Health Organization (WHO)<sup>2</sup>. This makes observations of weight, length, and weight-for-length attractive as measures of a child's long- and short-term health<sup>3,4</sup>. One of the main challenges to analysing child growth data is that individual growth trajectories display highly variable and complicated dynamic behaviour, differing markedly between children, even from the same geographic and socio-economic group. As such, developing growth models from which actionable insights can be extracted – such as identification of interventions at the population level or predictive risk assessments at individual child level – is both methodologically and practically challenging. Here we introduce a new methodology, stochastic differential equation (SDE)<sup>5</sup> models, into child growth research.

SDEs describe highly flexible dynamic processes comprising of two components: drift – gradual smooth changes, which could reflect developmental biological aspects such as physiology, nutrition or gut microbiome<sup>6</sup>; and diffusion – sudden short-term perturbations or shocks – like seasonal food insecurity, illness<sup>7</sup> or infection<sup>8</sup>. This stochastic behaviour could potentially help explain the large variability seen in growth trajectories. SDEs are extensively used in certain specialised applications, most notably in financial modelling<sup>9,10</sup>, to cope with the complicated dynamics of stock price movements. Some case studies utilizing SDEs exist in medicine and biology<sup>11</sup>, but they are not yet a part of a typical epidemiologist's or statistician's modelling toolbox.

To date, a wide range of different statistical curve-fitting methodologies have been applied to child growth trajectories, from common classical approaches such as hierarchical linear mixed models<sup>12</sup>, through to methods such as linear spline multilevel/broken stick<sup>13</sup> models, SITAR<sup>14</sup> growth curves, dynamic regression models<sup>15</sup> and functional principle component models<sup>16</sup>. SDEs are not curve fitting models but continuous time stochastic processes capable of rich dynamic behaviour. Given this, SDEs can enhance mechanistic interpretation of the drivers of variability (both long- and short-term) in growth in addition to improving forecasts of growth by more accurately capturing external sources of variability and uncertainty.

In the Methods section we provide a brief overview of SDE models. We present a minimum of theory, using instead two empirical case studies to introduce the key features of SDE modelling and how it can be readily applied in practice. In the Results section we present a more complex case study, including quantifying the impact of covariates on growth, using data from the two African sites of the MAL-ED study<sup>17</sup>. We conclude with a brief discussion of the opportunities for the application of SDEs in child growth research and outline some existing challenges. The computer code required to repeat the modelling results presented are provided as *Extended data*<sup>18</sup>.

## Methods

### Data sets and initial exploration

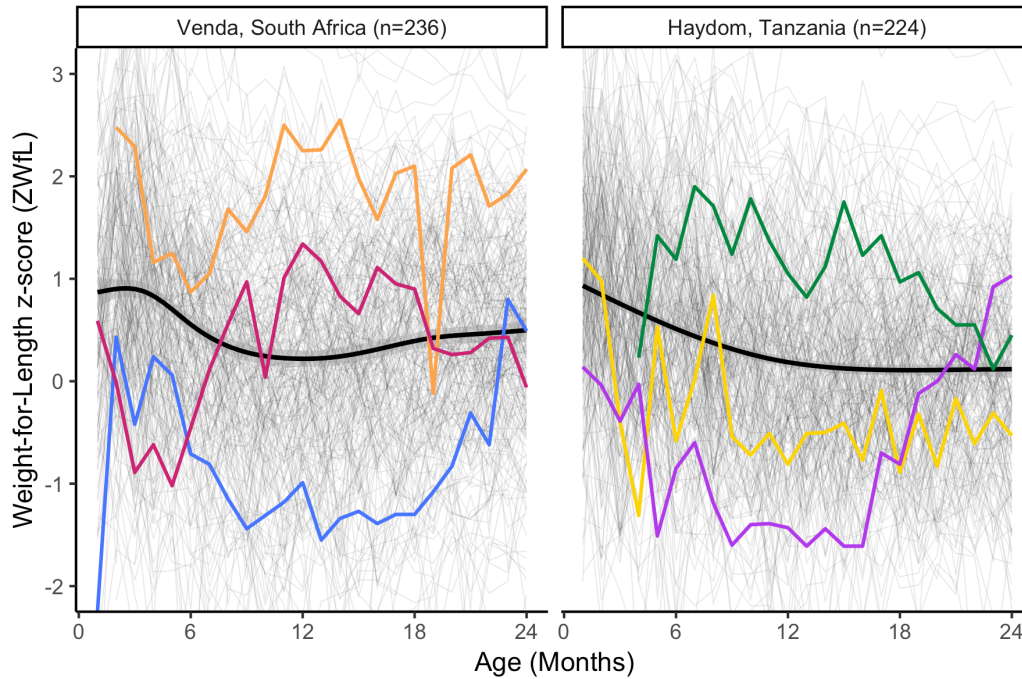
We use individual child data from the MAL-ED study, whose protocols, methodology and aggregate growth results have been presented previously<sup>8,17</sup>. MAL-ED was initiated as a multi-country cohort study located across eight low- and middle-income sites with historically high incidence of diarrhoeal disease and undernutrition, with a research focus on investigating determinants of development in children from birth through early years. Ethical approval was obtained from the National Institute of Medical Research for Tanzania (NIMR/HQ/R.8a/Vol.IX/858 and NIMR/HQ/R.8c/Vol.II/1034) and University of Venda Ethics Committee for South Africa (SMNS/09/MBY/004). Approval was additionally given by the institutional review board of the University of Virginia, USA and all methods used in this study followed the relevant guidelines and regulations. Informed consent was taken from parents of all children prior to enrolment.

In our case studies we use data from Haydom, Tanzania<sup>19</sup> (n = 224) and Venda, South Africa<sup>20</sup> (n = 236). Our focus here is on anthropometric data from ages 0–24 months where each child included in these analyses had between 20 and 25 monthly observations (within a window of  $\pm 14$  days), with 83% of children in Haydom and 86% in Venda having at least 24 observations. Weight and length, collected by trained fieldworkers and with minimal measurement error<sup>8</sup>, were converted to age and sex standardised z-scores using the WHO growth standards<sup>21</sup>. Here we focus on weight-for-length data, reflective of the relative weight of a child given their stature and therefore a child's current nutritional status<sup>22</sup>, and one of the growth z-scores recommended by the WHO for diagnosing acute malnutrition<sup>23</sup>. [Figure 1](#) shows all trajectories for weight-for-length z-score (ZWfL), along with the site-specific cross-sectional means.

High within- and between-child variability is the predominant feature of the raw trajectory data, which holds from birth through 24 months and for both sites. Three trajectories are highlighted in each site, and these illustrate the dynamic complexity of each child's growth.

### SDE models

The standard introductory text for SDEs is by Øksendal<sup>5</sup>, which contains a detailed mathematical exposition of SDEs. We focus on application and SDE models are introduced through examples with technical details largely omitted. We begin with a well-studied special case SDE model which we fit to data from three



**Figure 1. Observed weight-for-length trajectories for children from Venda and Haydom, 0–24 months, each thin grey line is an individual child and the thick black line is the population mean.** Large within and between child variability is clear, with three random children highlighted in each site.

individual children (separately) and compare results with a linear regression model. We then introduce a more general SDE model and fit this to data from all children from the Haydom site.

**Example 1 – Ornstein-Uhlenbeck (OU) process for a single trajectory.** Our growth outcome of interest is ZWfL, where the first observation in the growth trajectory for a single child is  $X_0$  when the child is aged  $t_0$ . We now develop a model to estimate likely values of  $X_1$ , ZWfL at a later age,  $t_1$ . In the simplest linear context, we could formulate an expression for the mean of  $X_1$  conditional on the previous observation,  $X_0$ , and the difference in age between  $t_1$  and  $t_0$ . Assuming  $X_1$  is approximately normally distributed with mean  $\mu$  and variance  $\sigma^2$ ,  $N(\mu, \sigma^2)$ , this would give a model of the form

$$X_1 | X_0 \sim N(X_0 + b(t_1 - t_0), \sigma^2), \tag{1}$$

which models ZWfL at a subsequent age as a linear function of the current value and the elapsed age. In this model, growth velocity – rate of change per unit time (age) – is described by  $b$ , a parameter to be estimated from the observed data. This is a fully specified growth model between any two time points, although too simple to be practically useful; firstly, the growth velocity,  $b$ , is assumed constant, and secondly, the variance  $\sigma^2$ , is constant and independent of the duration of the elapsed age between  $t_1$  and  $t_0$ , whereas it might reasonably be expected that two time points closer together in age may be more similar than those further apart.

Consider the same example as above but now where we have

$$X_1 | X_0 \sim N\left(e^{(t_1 - t_0)\alpha}(X_0 - \beta) + \beta, \frac{e^{-2t_1\alpha}(e^{2t_1\alpha} - e^{2t_0\alpha})\sigma^2}{2\alpha}\right). \tag{2}$$

This model now has three parameters,  $\alpha$ ,  $\beta$  and  $\sigma$ , and (2) is the transitional probability density function (or slice density) for the OU stochastic process. The OU model is well studied and has applications in mathematical finance<sup>24</sup> and theoretical biology<sup>25</sup>. The distribution in (2) is a more flexible model compared to (1), with non-linear mean and variance terms, both of which depend on the elapsed time between  $t_1$  and  $t_0$ . The OU process is typically defined in differential form by the stochastic differential equation

$$dX_t = \alpha(\beta - X_t)dt + \sigma dW_t, \tag{3}$$

where  $\alpha(\beta - X_t)$  is called the drift,  $\sigma$  is the diffusion and  $W_t$  is a Wiener (Brownian motion) process. The drift can be thought of as the slow-moving trend in growth velocity, while the diffusion is the continual perturbation of the system giving rise to volatility in velocity and therefore growth. This SDE has correlated movements through time, for example, in an OU process that commenced at time  $t_0$  the covariance between any two points in time,  $t$  and  $s$ , is  $(e^{-(s+t)\alpha}(-e^{2t_0\alpha} + e^{2\alpha \text{Min}\{s,t\}})\sigma^2)/(2\alpha)$ . Equation (3) provides the interpretation of the model parameters as components of the rate of change of growth, where (2) can be derived from (3) by solving the Fokker-Planck equation (see later).

To demonstrate the practical application of an SDE model to real data, we fitted OU models to three different ZWfL trajectories from the Haydom data. We compared these with the fit of a classical linear regression (LR) model, e.g.  $ZWfL(t) \sim N(a_0 + a_1t, \sigma^2)$ , with both models having three parameters. Figure 2 shows the raw trajectory data, along with fitted values from the OU and LR models (separately for each child). See SI.1 (Extended data)<sup>18</sup> for the model fitting code. In the OU model each successive data point (say,  $ZWfL_{t_1}$ ) depends on the previous point ( $ZWfL_{t_0}$ ), which means the OU model fit is not a smooth curve, but transitions from point to point. In the LR model, the fitted value at age  $t_1$  does not depend directly on the previously observed value  $ZWfL_{t_0}$ , but rather it is from a globally parameterized smooth curve computed across all observations. The OU model is a special case of a more general SDE formulation, which we present next, along with the key concepts in fitting SDEs to data.

**Defining an SDE model – basic concepts.** A general formulation for an SDE model (see 7.1.2 in Øksendal<sup>5</sup>) is

$$dX_t = \mu(t, X_t, \theta)dt + g(t, X_t, \theta)dW_t, \tag{4}$$

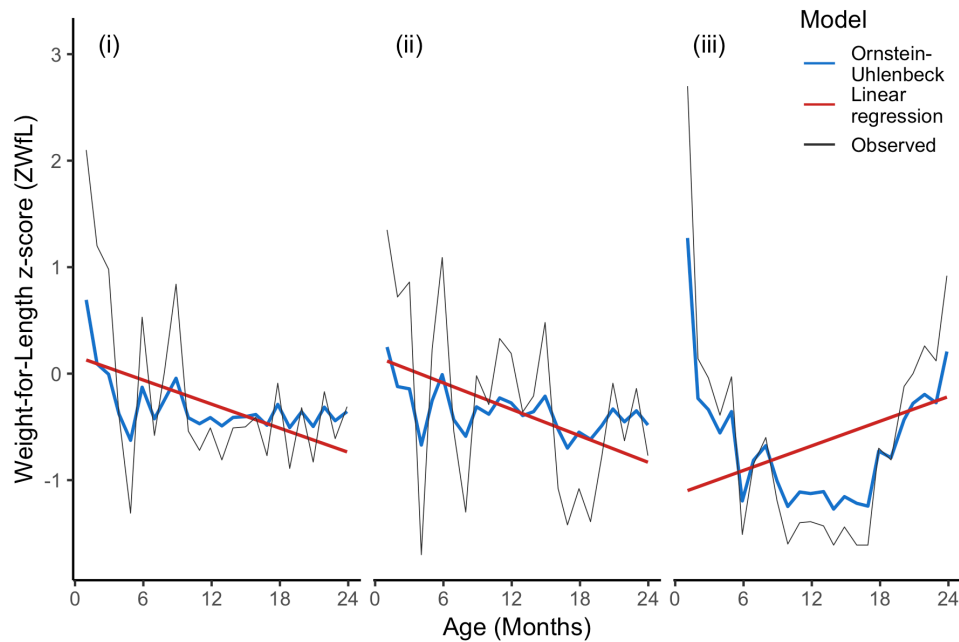
where  $X_t$  is our growth outcome variable of interest (e.g. ZWfL). Functions  $\mu(t, X_t, \theta)$  and  $g(t, X_t, \theta)$  generalise the drift and diffusion terms from (3), respectively, where  $\theta$  is a set of parameters to be estimated and, as before,  $W_t$  is a Wiener process. The OU model from (3) is therefore a special case of (4) where  $\mu(t, X_t, \theta) = \alpha(\beta - X_t)$  and  $g(t, X_t, \theta) = \sigma$ . Equation (4) can also be written as a stochastic integral equation

$$X_t = X_0 + \int_0^t \mu(s, X_s, \theta)ds + \int_0^t g(s, X_s, \theta)dW_s, \tag{5}$$

which usefully emphasises that these are models of the evolution of a continuous time stochastic process – here growth of a child. Equation (5) says that the value of growth (a random variable) for a child at  $t$  time units into the future from the currently observed time ( $t=0$ ) is the current value of growth ( $X_0$ ) plus the sum over this time interval of the drift and diffusion components of velocity. As the diffusion is integrated with respect to a “white noise” process  $W_s$ , then  $X_t$  follows a probability distribution at time  $t$ . To make this more concrete, if  $ZWfL_{t_1}$  is our growth outcome at time  $t_1$ , then  $f(ZWfL_{t_1})$  is the probability distribution of  $ZWfL_{t_1}$  conditional on  $ZWfL_{t_0}$ , the previously observed value of growth at time  $t_0$ . This function,  $f(X_t)$ , which is essential for model fitting, is a solution to the partial differential equation

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial x} \left( \frac{1}{2} \frac{\partial (g^2(t, X_t, \theta)f)}{\partial x} - \mu(t, X_t, \theta)f \right). \tag{6}$$

Equation (6) is called the Fokker-Planck or forward-Kolmogorov equation (whose complete specification includes initial and boundary conditions, which have been omitted). The SDE in equation (4) defines the terms in Equation (6). Solving Equation (6) gives the expression for  $f(X_t)$ . If we consider again the OU process and plug in the relevant terms from (3) into (6) and do the necessary calculus, then we arrive at exactly the normally distributed slice density stated in (2).



**Figure 2. Comparison of stochastic differential equation (SDE assuming an Ornstein-Uhlenbeck process) model and linear regression model for three children from Haydom, where both models have three parameters.** The difference between a curve fitting approach (fit globally) and time series approach (based on modelling the change over time) is clearly evident.

For model fitting we need to compute the likelihood function given the slice density. If we consider first the likelihood function for trajectory data from a single child, and where we have  $N$  observations over time, then the negative log-likelihood for a single child can be written as (see Hurn<sup>10</sup>)

$$-\log \mathcal{L}(\theta) = -\log f_0(X_0 | \theta) - \sum_{k=0}^{N-1} \log f(X_{k+1} | X_k, \theta), \quad (7)$$

where  $f_0(X_0|\theta)$  is the probability density of the growth outcome variable at the first available data point,  $f(X_{k+1} | X_k, \theta) = f(X_{k+1}, t_{k+1} | (X_k, t_k), \theta)$  is the value of the slice density function for a stochastic process starting at  $(X_k, t_k)$  and evolving to  $(X_{k+1}, t_{k+1})$ . Equation (7) allows us to compute the likelihood function for all trajectories, including, if necessary, covariance structures across children through the inclusion of random effects (see later). One minor remark is how to deal with the first available observation,  $X_0$ , as SDE models are defined in terms of transitions. We follow Schneider<sup>26</sup> and the existing literature in maximum likelihood estimation in SDEs and treat  $X_0$  as a constant, which is also typical in the time series literature.

In summary, the key steps for working with SDE models focussed on model fitting are: (i) choose a form of  $\mu(t, X_t, \theta)$  and  $g(t, X_t, \theta)$  in (4) and then; (ii) determine the corresponding slice density which satisfies the Fokker-Planck Equation (6) and then; (iii) fit the model to the data using the likelihood function in (7). Here we only consider forms of  $\mu(t, X_t, \theta)$  and  $g(t, X_t, \theta)$  which have known analytical solutions (slice densities) to (6), which then makes fitting SDE models to data no different from a standard maximum likelihood problem using standard statistical software. Mathematica software (version 11.3, Wolfram Research Inc.), for example, can be used to compute slice densities for a wide range of SDEs, and a selection of these solutions is provided for reference in SI.2 (see *Extended data*)<sup>18</sup> as illustration. While in theory it is possible to fit SDEs that do not have an analytical solution to (6) to data, in practice this is numerically challenging (see the Discussion).

**Example 2 - OU process for multiple trajectories.** In the Haydom data we have 224 trajectories (children) across 0–24 months, and we now fit (non-linear) mixed model variants of the OU process, along with standard linear mixed models to these data. Table 1 gives a summary of different parameterizations and goodness of fit metrics. The modelling code is provided in SI.3a along with model output SI.3b (see *Extended data*)<sup>18</sup>.

For the same number of parameters, the OU process gives substantially better Akaike information criterion (AIC) and Bayesian information criterion (BIC) metrics, and fitting these SDE models including random effects is straightforward, requiring only with a few lines of code in SAS’s proc nlmixed. These mixed models can also be implemented in the Stan<sup>27</sup> language, with an OU specific example using Stan provided by Goodman (2018)<sup>28</sup>.

**Main case study - model formulations**

Our main results comprise of an illustrative case study where we considered the combined data from Haydom and Venda. The general model formulations considered, and model search process are detailed below. To keep the analysis as clear as possible we considered only one covariate, (in addition to age) in the modelling, a categorical variable indicating site.

**Linear mixed models (LMMs).** We considered LMMs where the most general formulation for ZWfL (response) was: a polynomial function of age (continuous) and site (two categories); with interactions between age and site; fixed and random (normally distributed) effects for the age terms in the polynomial; where random effects for the age terms (including intercept) had an unstructured covariance matrix; and within child errors had an AR(1) – autoregressive first order – covariance structure to allow serial dependence between errors. Increasing orders of polynomial (up to fourth order) were examined, guided by AIC and BIC metrics, in addition to trimming terms with high p-values (>0.1). These model formulations can be readily fitted to trajectory data using proc mixed in

**Table 1. Model goodness of fit comparisons using a selection of stochastic differential equation and linear regression models.**

Model	No. parameters	Remarks	AIC (smaller is better)	BIC (smaller is better)
Ornstein-Uhlenbeck (Equation 3)	3	No random effects	<b>12181</b>	<b>12201</b>
Linear regression			15454	15473
Ornstein-Uhlenbeck with mixed effects	6	Random speed of reversion <sup>1</sup> and long-term mean with covariance <sup>2</sup>	<b>11544</b>	<b>11564</b>
Linear mixed effects regression		Random intercept and slope with covariance <sup>2</sup>	11607	11627

AIC, Akaike information criterion; BIC, Bayesian information criterion. <sup>1</sup> $\alpha$  in the Wiener (Brownian motion) process and is the covariance between time points (see text above for detail); <sup>2</sup>Child-level random effects.

SAS. Relevant SAS code is provided in SI.4a (see *Extended data*)<sup>18</sup>. This can also be achieved using Stan, for example adapting Goodman (2018)<sup>28</sup>.

**SDE (non-linear) mixed models.** The most general formulation considered for the SDE models was

$$dX_t = (\alpha_2 + \alpha_3 t + \alpha_4 t^2 + \alpha_5 t^3 + \alpha_1 X_t)dt + \sigma dW_t, \quad (8)$$

where  $X_t$  is ZWfL at age  $t$  and so the rate-of-change per unit time for ZWfL depends on both the current age of the child and the child’s current value of ZWfL. Specifically, we considered a polynomial (a cubic in equation 8) function of age, plus a linear term ( $\alpha_1 X_t$ ) in ZWfL. This polynomial formulation of SDE in (8) has a closed form of slice density (a normal distribution) if we hold  $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \sigma)$  constant. Mathematica was used to determine the slice density which is

$$X_t | X_0 \sim N \left( \frac{1}{\alpha_1^2} e^{-t\alpha_1} (e^{t\alpha_1} x_0 \alpha_1^4 - 6(e^{t\alpha_1} - e^{t_0\alpha_1})\alpha_5 + \alpha_1(-2(e^{t_0\alpha_1} - e^{t\alpha_1})\alpha_4 - 6(-e^{t\alpha_1} t_0 + e^{t_0\alpha_1} t_1)\alpha_3) + \alpha_1^2((-e^{t_0\alpha_1} + e^{t\alpha_1})\alpha_3 + 2e^{t\alpha_1} t_0 \alpha_4 + 3e^{t_0\alpha_1} t_0^2 \alpha_5 - e^{t_0\alpha_1} t_1(2\alpha_4 + 3t_1\alpha_5)) + \alpha_1^3((-e^{t_0\alpha_1} + e^{t\alpha_1})\alpha_2 + e^{t\alpha_1} t_0 \alpha_3 + e^{t_0\alpha_1} t_0^2 \alpha_4 + e^{t\alpha_1} t_0^3 \alpha_5 - e^{t_0\alpha_1} t_1(\alpha_3 + t_1(\alpha_4 + t_1\alpha_5))))), \frac{(-1 + e^{2(-t_0+t_1)\alpha_1})\sigma^2}{2\alpha_1} \right). \quad (9)$$

Comparing Equation (8) with (9) demonstrates how compact the differential form of the model is, but it is the slice density which is required for model fitting and parameter estimation. The most general form of (8) we considered additionally allowed each of  $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \sigma)$  to be included in the model as both a fixed and random (normally distributed) effect, thereby allowing trajectories to be tailored to individual children. An unstructured covariance matrix for the random effects with simplifications down to a diagonal covariance matrix were considered. The model search considered increasing orders of polynomial up to cubic and in keeping with the LMM search was guided by AIC and BIC metrics, in addition to trimming terms with high p-values (>0.1). Relevant SAS code is provided in SI.4a (see *Extended data*)<sup>18</sup>. This can also be achieved using Stan, for example, adapting Goodman (2018)<sup>28</sup>.

**Prediction**

The main real-world application area of SDEs is in predictive modelling (e.g. Iversen *et al.*<sup>26</sup>). Here we use SDE models to predict future growth given a child’s current age and current ZWfL. Such predictions have two application areas: 1) as part of an individual child’s risk assessment to determine if they require an intervention; and 2) to elucidate structure hidden within the large variability across growth trajectories, which may then offer new insights into drivers of growth at the population level. We use our best fitting SDE model to predict future growth trajectories across a grid of starting points for age and ZWfL, separately for each of the two sites.

Predictions are calculated using a 10-fold random sampling approach, where we draw from all the parameters estimated (fixed and random) in the best fitting SDE model – one set for each child. Which parameter sets are chosen to generate predictions

depends on how likely trajectories generated from each set are to have visited each given starting point across an age-ZWfL grid. This adds an important element of “locality” to our predictions, combined with 10-fold sampling to provide an indication of robustness of our predictions. A more detailed description of the prediction algorithm is given below, with full R code provided in SI.5 (see *Extended data*)<sup>18</sup>.

For each point across an age-ZWfL grid we compute the likelihood of observing this point for each set of parameters, using the relevant slice density, where the initial starting point is the first age available for each trajectory. Predictions from each grid point progress through increasing ages using the new slice distribution at each next point in time. For example, for predictions in Venda we have n=236 likelihood values for each age-ZWfL grid point. The most likely parameter set is then used for the prediction of the next ZWfL at the next age, with 10-fold sampling used to indicate how robust this prediction is. The 10-fold sampling splits these n=236 parameter sets into 10 random groups, and within each group we choose the parameter set with the highest likelihood value as the one to be used for the next prediction. In summary, from a fixed starting point in an age-ZWfL grid we have a main prediction at each future age up to 24 months, plus 10 additional predictions at each age as a sensitivity analysis (incrementing age in small steps).

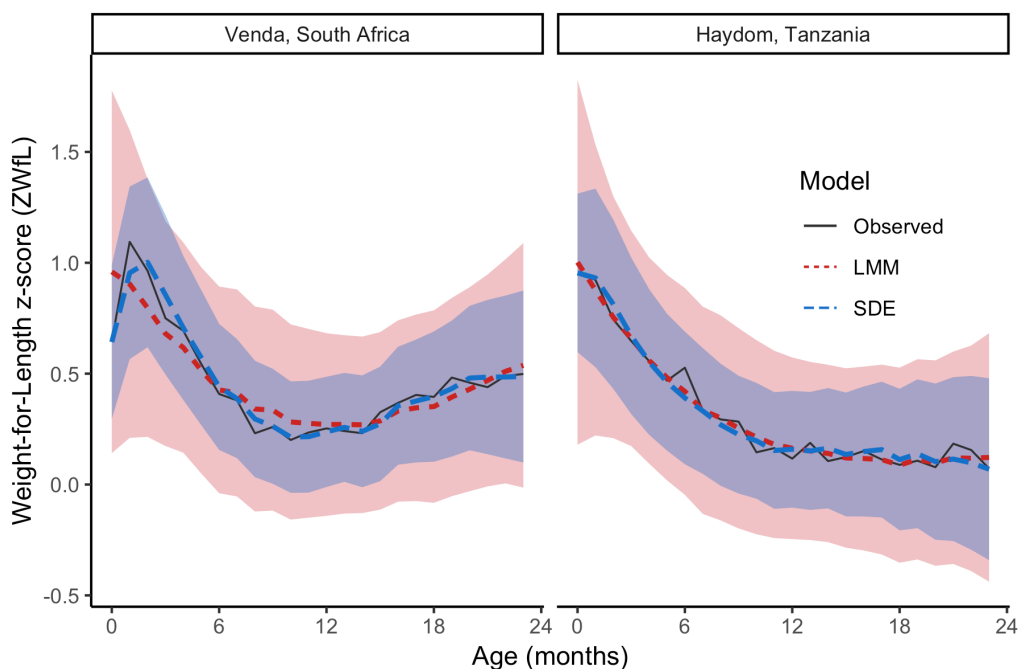
**Results**

**Linear mixed models and SDE mixed models**

Using individual child trajectory data over 0–24 months for ZWfL from all n=460 children from the two sites, the best fitting LMM model was a cubic polynomial with a single interaction term between age squared and site (with no separate term for site), which gave AIC=21913 and BIC=21983. The best fitting SDE model had corresponding values of AIC=21718 and BIC=21772, where this model had random effects in four of the six model parameters, a diagonal covariance matrix and site included in two of the drift parameters and the diffusion parameter. Full modelling descriptions can be found in SI.4a and results, including parameter estimates, can be found in the SI.4.b (see *Extended data*)<sup>18</sup>. The AIC and BIC metrics suggest the SDE offers an improved fit to the data. Examining residuals and fitted values the fit of each model is visually similar; however, there are notable qualitative and quantitative differences.

Figure 3 shows observed monthly means compared with estimates of the population mean ZWfL from each model. These are quantitatively similar except that the SDE correctly captures the initial shape of the mean, a short rapid increase then decrease during the first six months in the Venda data, whereas the LMM estimates a steady decrease from birth through six months. Figure 3 also shows confidence intervals for the population mean of ZWfL from each model over age; the SDE model has considerably narrower intervals suggesting it has explained more of the variation in the data than the LMM model.

In summary, our results so far suggest that our SDE model is at least as good, and appears superior in some respects, to a reasonable choice of classical LMM.



**Figure 3. Comparison between population means estimated from stochastic differential equation (SDE) and linear mixed models (LMM) with 95% confidence intervals, and observed monthly means.** SDE model has narrower confidence intervals in each site, entirely contained within the wider LMM confidence interval.

### Prediction

Figure 4 shows predictions of the most likely future ZWfL trajectory for a child, conditional on current age and ZWfL values. A grid of starting values for age and ZWfL within the ranges observed (from both sites) was used, with separate predictions for each site. There are several particularly striking features in these predictions; most notably, they suggest the presence of longitudinal clustering within each site, as we find what appear to be a small distinct set of “paths” or “streams” hidden inside the large between-child variability in trajectories observed in Figure 1. These clusters also appear to differ quantitatively and qualitatively between sites, with fewer clusters in Haydom and of different shapes to those in Venda. Further, these clusters imply some degree of canalisation of trajectories, for example, the small number of children who are wasted with ZWfL values of  $\leq -2$  (Venda,  $n = 16/236$ ; Haydom,  $n = 2/224$ ) converge on similarly low predicted values (the rapid increase from extreme values [e.g.  $\leq -3$ ] shows that predictions return to values where more data are present).

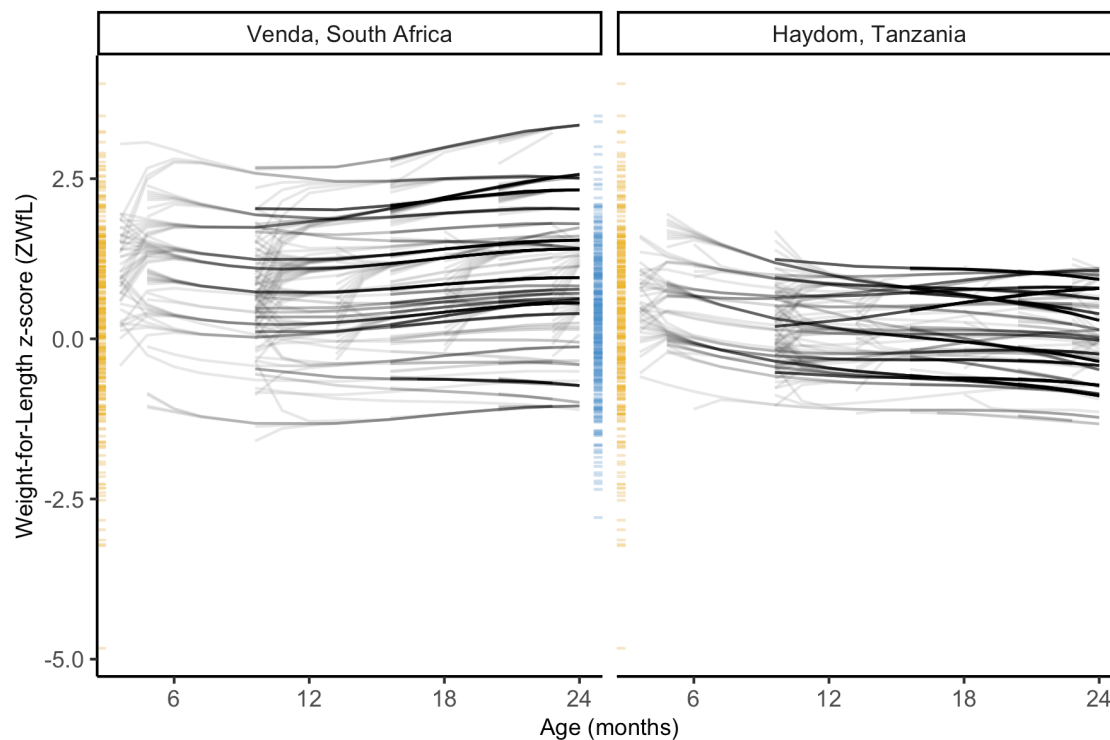
### Discussion

We have presented a novel approach for analysing child growth trajectories, using a modelling methodology, stochastic differential equations, widely used in other fields but not yet in child growth research. The use of a continuous time stochastic

process approach, such as SDEs, to model child growth trajectories is conceptually appealing as it explicitly acknowledges - through drift and diffusion processes - the highly complicated dynamic environment into which a new born child is delivered and exposed, particularly in resource-limited settings. Our results show that SDEs also have practical appeal as they offer very different (highly non-linear) formulations from the usual additive linear models, which gave good results with our case study data. This suggests that SDEs may be an attractive alternative to other established methods, at least as supporting analyses, moreover because SDEs can also be readily fitted using standard software such as SAS or open source alternatives such as Stan. Our supplementary information contains modelling code that can be adapted to other study data<sup>18</sup> or as a basis for comparison to other methods and alternative drift/diffusion terms.

While only an initial exploration of a subset of the MAL-ED data using SDE modelling, our predictive results presented in Figure 4 were both unexpected and exciting. These results suggest that the predictive capability of SDE models could potentially reveal new insights hidden by the large between-child variations that typify child growth. For example, here the prediction of ZWfL implies canalisation of growth trajectories, with particular implications for children who start life wasted (or close to) and are predicted to remain so through the first two years of life, although it is worth noting that observations





**Figure 4. Predicted future weight-for-length trajectories from a grid of starting points using an SDE model.** Yellow ticks indicate the observed weight-for-length at 0 months; blue ticks indicate the observed weight-for-length at 24 months. The best-fitting predictions are shown in red and 10-fold cross validation in grey. The distributions of observed weight-for-length at both zero and 24 months in each site is shown to the right of the respective plot. The predictions appear to cluster together into a smaller set of “paths”, which also differ between sites.

of wasting were rare in these two populations. The predictive method presented is relatively ad-hoc and simple – prediction in SDEs with random effects is novel - and is an area in need of development.

We restricted our presentation to a narrow selection of simple SDE models, many more parameterizations are available with explicit expressions for the slice density (e.g. using software like Mathematica). More complex formulations, particularly for the diffusion function  $g(t, X_t, \theta)$ , may add considerable richness to an SDE model’s dynamic behaviour, for example, by incorporating additional covariates such as seasonality or food security, morbidity or even measurement error; however, these would require numerical methods to compute the likelihood function. Initial explorations suggest this is far from straightforward, both in terms of computational feasibility and also numerical stability, and is another area ripe for future development.

### Data availability

#### Underlying data

Data from the MAL-ED study are available from <https://clinepidb.org/>. Guest users can view data and access analysis tools and record pages, but must obtain approval from the data

providers to download data. The request may be submitted via a form that pops up when a user logs in with a registered account and clicks “Download data”.

### Extended data

Zenodo: Introducing a drift and diffusion framework for childhood growth research. <https://doi.org/10.5281/zenodo.4278002><sup>18</sup>.

This project contains the following extended data:

- SI1.pdf (Model code for OU and linear regression models, in SAS)
- SI2.pdf (Illustrative reference slide densities, using Mathematica)
- SI3a.pdf (SAS code to compare OU and LMM models for Table 1)
- SI3b.pdf (SAS model output comparing OU and LMM models in Table 1)
- SI4a.pdf (SAS code to fit mixed effects OU and LMM models)
- SI4b.pdf (SAS model output for OU and LMM models)

- SI5.pdf (R code to generate figures, including the prediction algorithm to generate Figure 4)

Data are available under the terms of the [Creative Commons Attribution 4.0 International license \(CC-BY 4.0\)](#).

#### Code availability

Source code available from: [https://github.com/fraseriainlewis/gatesopenresearch\\_SI](https://github.com/fraseriainlewis/gatesopenresearch_SI)

Archived code at time of publication: <https://doi.org/10.5281/zenodo.4278002><sup>18</sup>

License: [Creative Commons Attribution 4.0 International license \(CC-BY 4.0\)](#).

#### Acknowledgements

We thank the participants and staff of the MAL-ED study for their vital contributions and we thank Prof. Laura Caulfield for her insightful and constructive input. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the U.S. National Institutes of Health or Department of Health and Human Services.

#### References

- Cole TJ: **The development of growth references and growth charts.** *Ann Hum Biol.* 2012; **39**(5): 382–394.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- WHO MULTICENTRE GROWTH REFERENCE STUDY GROUP, de Onis M: **WHO Child Growth Standards based on length/height, weight and age.** *Acta Paediatrica.* 2006; **95**(s450): 76–85.  
[Publisher Full Text](#)
- Black RE, Allen LH, Bhutta ZA, *et al.*: **Maternal and child undernutrition: global and regional exposures and health consequences.** *Lancet.* 2008; **371**(9608): 243–260.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Victora CG, Adair L, Fall C, *et al.*: **Maternal and child undernutrition: consequences for adult health and human capital.** *Lancet.* 2008; **371**(9609): 340–357.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Øksendal B: **Stochastic Differential Equations: An Introduction with Applications.** (Springer), 2007.  
[Reference Source](#)
- Yatsunenkov T, Rey FE, Manary MJ, *et al.*: **Human gut microbiome viewed across age and geography.** *Nature.* 2012; **486**(7402): 222–227.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mata LJ: **The Children of Santa Maria Cauqué: A Prospective Field Study of Health and Growth.** 1978.  
[Reference Source](#)
- The MAL-ED Network Investigators: **Relationship between growth and illness, enteropathogens and dietary intakes in the first 2 years of life: findings from the MAL-ED birth cohort study.** *BMJ Global Health.* 2017; **2**(4): e000370.  
[PubMed Abstract](#) | [Free Full Text](#)
- Lux T: **Rational forecasts or social opinion dynamics? Identification of interaction effects in a business climate survey.** *J Econ Behav Organ.* 2009; **72**(2): 638–655.  
[Publisher Full Text](#)
- Hurn AS, Jeisman JJ, Lindsay KA: **Seeing the Wood for the Trees: A Critical Evaluation of Methods to Estimate the Parameters of Stochastic Differential Equations.** *J Financ Economet.* 2007; **5**(3): 390–455.  
[Publisher Full Text](#)
- Allen LJS: **A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis.** *Infect Dis Model.* 2017; **2**(2): 128–142.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Burchinal MR, Bailey DB, Snyder P: **Using Growth Curve Analysis to Evaluate Child Change in Longitudinal Investigations.** *J Early Intervention.* 1994; **18**(4): 403–423.  
[Publisher Full Text](#)
- Howe LD, Tilling K, Matijasevich A, *et al.*: **Linear spline multilevel models for summarising childhood growth trajectories: A guide to their application using examples from five birth cohorts.** *Stat Methods Med Res.* 2013; **25**(5): 1854–1874.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cole TJ, Donaldson MDC, Ben-Shlomo Y: **SITAR—a useful instrument for growth curve analysis.** *Int J Epidemiol.* 2010; **39**(6): 1558–1566.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ivanescu AE, Crainiceanu CM, Checkley W: **Dynamic child growth prediction: A comparative methods approach.** *Stat Model.* 2017; **17**(6): 468–493.  
[Publisher Full Text](#)
- Han K, Hadjipantelis PZ, Wang JL, *et al.*: **Functional principal component analysis for identifying multivariate patterns and archetypes of growth, and their association with long-term cognitive development.** *PLoS One.* 2018; **13**(11): e0207073.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- MAL-ED Network Investigators: **The MAL-ED Study: A Multinational and Multidisciplinary Approach to Understand the Relationship Between Enteric Pathogens, Malnutrition, Gut Physiology, Physical Growth, Cognitive Development, and Immune Responses in Infants and Children Up to 2 Years of Age in Resource-Poor Environments.** *Clin Infect Dis.* 2014; **59** Suppl 4: S193–S330.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- McCormick B: **Introducing a drift and diffusion framework for childhood growth research (Version 1).** *Zenodo.* 2020.  
<http://www.doi.org/10.5281/zenodo.4278002>
- Mduma ER, Gratz J, Patil C, *et al.*: **The Etiology, Risk Factors, and Interactions of Enteric Infections and Malnutrition and the Consequences for Child Health and Development Study (MAL-ED): Description of the Tanzanian Site.** *Clin Infect Dis.* 2014; **59**(Suppl 4): S325–S330.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bessong PO, Nyathi E, Mahopo TC, *et al.*: **Development of the Dzimauli Community in Vhembe District, Limpopo Province of South Africa, for the MAL-ED Cohort Study.** *Clin Infect Dis.* 2014; **59**(Suppl 4): S317–S324.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- WHO: **WHO Child Growth Standards: Methods and development: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age.** (WHO), 2006.  
[Reference Source](#)
- Waterlow JC: **Classification and definition of protein-calorie malnutrition.** *Br Med J.* 1972; **3**(5826): 566–569.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mwangome MK, Berkley JA: **The reliability of weight-for-length/height Z scores in children.** *Matern Child Nutr.* 2014; **10**(4): 474–480.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Nicolato E, Venardos E: **Option Pricing in Stochastic Volatility Models of the Ornstein-Uhlenbeck type.** *Math Financ.* 2003; **13**(4): 445–466.  
[Publisher Full Text](#)
- Bartoszek K, Glémin S, Kaj I, *et al.*: **Using the Ornstein-Uhlenbeck process to model the evolution of interacting populations.** *J Theor Biol.* 2017; **429**: 35–45.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Iversen EB, Morales JM, Møller JK, *et al.*: **Leveraging stochastic differential equations for probabilistic forecasting of wind power using a dynamic power curve.** *Wind Energy.* 2017; **20**(1): 33–44.  
[Publisher Full Text](#)
- Carpenter B, Gelman A, Hoffman MD, *et al.*: **Stan: A probabilistic programming language.** *J Stat Softw.* 2017; **76**(1).  
[Publisher Full Text](#)
- Goodman A: **Fitting Ornstein-Uhlenbeck-type Student's t-processes in Stan with applications for population dynamics data.** *Zenodo.* 2018.  
[Publisher Full Text](#)

## Open Peer Review

Current Peer Review Status:  

---

### Version 2

Reviewer Report 07 January 2021

<https://doi.org/10.21956/gatesopenres.14421.r30075>

© 2021 Cameron N. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



#### Noel Cameron

School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough, UK

I think that the submission can go forward because it is scientifically valid although I still am not convinced that this approach is useful or that it makes a significant contribution to what is already known. I don't think this modelling does provide greater insight into anthropometric data and don't think that authors have made a case for its significance over other methods. I remain to be convinced that it will be widely used in the analysis of growth data.

**Competing Interests:** No competing interests were disclosed.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 03 December 2020

<https://doi.org/10.21956/gatesopenres.14421.r30076>

© 2020 Berkley J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



#### James A. Berkley

<sup>1</sup> The Childhood Acute Illness and Nutrition (CHAIN) Network, Nairobi, Kenya

<sup>2</sup> University of Oxford, Oxford, UK

<sup>3</sup> KEMRI/Wellcome Trust Research Programme, Kilifi, Kenya

Thank you, my comments have been well-addressed and I believe the manuscript is suitable for indexing.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Child survival, infectious diseases, malnutrition, newborn health, clinical trials, anthropometry

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

## Version 1

Reviewer Report 10 August 2020

<https://doi.org/10.21956/gatesopenres.14305.r29206>

© 2020 Berkley J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**James A. Berkley** 

<sup>1</sup> The Childhood Acute Illness and Nutrition (CHAIN) Network, Nairobi, Kenya

<sup>2</sup> University of Oxford, Oxford, UK

<sup>3</sup> KEMRI/Wellcome Trust Research Programme, Kilifi, Kenya

This is an interesting and novel methods paper outlining an approach to analyzing child growth through translation of the SDE approach from physics and financial modelling. SDEs extend the concept common to growth modelling methods that individuals vary stochastically about a population curve.

- It would be helpful in the introduction to outline which uses the SDE approach is aiming at. For example, descriptive, explanatory or forecasting in populations or real-time use in individuals.
- In the introduction, for clarity please also include whether 'drift' is also thought to represent exposures such as seasonal food security and dietary intake patterns and feeding practices that may change with age (e.g. breastfeeding) which may be more influential than the described 'developmental biological aspects'.
- Reference 8 suggests that good reliability weight and length can be assumed, however length is more difficult to measure precisely in younger infants and I notice extreme values appear more common close to the time of birth on the figures, including 4 of the 6 individuals presented in figures 1 and 2. Translation to Z scores (a ratio) may compound errors. Was reliability assessed for (computer calculated) ZWFL? If the data are available, I suggest including it in this report, including performance across the age range 0-24m. It would also be worth commenting that apparent weight-for-length may be affected by hydration status as both may provide additional stochastic variability that varies with age.

The former may potentially affect the constant term  $X_0$  and the starting points across the age-ZWfL grid.

- An issue worth considering is that when examining changes in Z scores is that the standard deviation (the denominator for ZWfL) varies across the range of length (see: [https://www.who.int/childgrowth/standards/wfl\\_girls\\_0\\_2\\_zscores.txt](https://www.who.int/childgrowth/standards/wfl_girls_0_2_zscores.txt)). This means that children with a constant deficit in weight-for-length as they grow longer/heavier may demonstrate changes in ZWfL. Leroy *et al.* (2015<sup>1</sup>) proposed the use of an absolute difference criterion (for HAZ in their case) as a more robust way to assess changes in nutritional status.
- The formulations of the SDE models themselves is outside my expertise.
- In Methods, example 2, it may be helpful to the readers to point out if child ID was a random effect term in the non-linear mixed model variants and explain 'random speed of reversion'.
- The longitudinal clustering is very interesting, with potential genetic and environmental interpretations. However, the finding would be strengthened by validation using another method such as simple probability estimates for ending up in one of the final 'bins' identified in the model to reassure readers this is not an artifact of the method (quantization?) or due to different error in the first measurements (as above).
- The discussion should expand on future research that the method leads to. Anthropometry is a practical proxy for health and nutritional status and validation of criteria for intervention are typically based on mortality risk in large populations. As a next stage of research, prediction of actual health events such as death, serious illness or impaired neurodevelopment would be valuable to mention. Additionally, it would be helpful to comment on the models' ability to include time-varying covariates such as seasonality, illness or food security shock, as well as preterm birth.
- The narrower CIs suggest the SDE model is a better approximation of the data than the linear mixed models. However, linear models may not be regarded as a current gold standard for growth models, and comparisons with others including latent growth models, SITAR, multilevel fractional polynomial models or penalised splines etc. This is a limitation of the paper and area for future work.

## References

1. Leroy JL, Ruel M, Habicht JP, Frongillo EA: Using height-for-age differences (HAD) instead of height-for-age z-scores (HAZ) for the meaningful measurement of population-level catch-up in linear growth in children less than 5 years of age. *BMC Pediatr.* 2015; **15**: 145 [PubMed Abstract](#) | [Publisher Full Text](#)

## Is the work clearly and accurately presented and does it cite the current literature?

Yes

## Is the study design appropriate and is the work technically sound?

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

I cannot comment. A qualified statistician is required.

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Child survival, infectious diseases, malnutrition, newborn health, clinical trials, anthropometry

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 17 Nov 2020

**Ben McCormick,**

**It would be helpful in the introduction to outline which uses the SDE approach is aiming at. For example, descriptive, explanatory or forecasting in populations or real-time use in individuals.**

We have amended the introduction to note that we believe this approach to be particularly useful in explanatory analysis of growth data ("SDEs can enhance mechanistic interpretation of the drivers of variability"). We note that in the discussion we do discuss the advantage SDEs offer to forecasting because they are process-based models rather than purely descriptive. Forecasts ought, assuming appropriate covariables, better account for variance.

**In the introduction, for clarity please also include whether 'drift' is also thought to represent exposures such as seasonal food security and dietary intake patterns and feeding practices that may change with age (e.g. breastfeeding) which may be more influential than the described 'developmental biological aspects'.**

We have added these to the text, but we note that the terms that are included in the drift and diffusion are specific to the conceptual model examined. Certainly, these terms could be included (in either or both components) and it is a matter of the appropriate technical details. This reflects the richness of the method, because such exposures can be explored mechanistically. Diet might conceivably impact both drift ('usual' diet) or diffusion (seasonal food insecurity).

**Reference 8 suggests that good reliability weight and length can be assumed, however length is more difficult to measure precisely in younger infants and I notice extreme values appear more common close to the time of birth on the figures, including 4 of the 6 individuals presented in figures 1 and 2. Translation to Z scores (a ratio) may compound errors. Was reliability assessed for (computer calculated) ZWfL? If the data are available, I suggest including it in this report, including performance across the age range 0-24m. It would also be worth commenting that apparent weight-for-length may be affected by hydration status as both may provide additional stochastic variability that varies with age. The former may potentially affect the constant term  $X_0$  and the starting points across the age-ZWfL grid.**

Reliability (noted in reference 8) was measured on the basis of 5% remeasurement. Reliability was indeed lower for measurements of younger children, and lower for the combined WFL than either weight or length, albeit with correlation coefficients over a range of 0.92 (1 month) to 0.98 (24 months). We have noted that this could be incorporated into the model.

**An issue worth considering is that when examining changes in Z scores is that the standard deviation (the denominator for ZWfL) varies across the range of length (see: [https://www.who.int/childgrowth/standards/wfl\\_girls\\_0\\_2\\_zscores.txt](https://www.who.int/childgrowth/standards/wfl_girls_0_2_zscores.txt)). This means that children with a constant deficit in weight-for-length as they grow longer/heavier may demonstrate changes in ZWfL. Leroy *et al.* (2015<sup>1</sup>) proposed the use of an absolute difference criterion (for HAZ in their case) as a more robust way to assess changes in nutritional status.**

The Leroy *et al.* argument is particularly interesting in the context of recovery from stunting, which tends to occur post-24 months. For this particular study, WFL was used as an example and the methodology would be equally applicable to other anthropometric metrics.

**The formulations of the SDE models themselves is outside my expertise.**

It is our hope that this introduction may inspire others to consider this approach and that it will make the methods less opaque by demonstrating their application and accessibility in conventional software.

**In Methods, example 2, it may be helpful to the readers to point out if child ID was a random effect term in the non-linear mixed model variants and explain 'random speed of reversion'.**

We have now noted that these were child-level parameters (random effects). The random speed of reversion is the covariance between time points, so the closer time points are the more similar they are. We have added a note to the text.

**The longitudinal clustering is very interesting, with potential genetic and environmental interpretations. However, the finding would be strengthened by validation using another method such as simple probability estimates for ending up in one of the final 'bins' identified in the model to reassure readers this is not an artifact of the method (quantization?) or due to different error in the first measurements (as above).**

We apologise that we mis-wrote the code for the prediction presented in figure 4. The model to evaluate predictions was correct, but the 10-fold selection of the best fitting points

incorrectly selected observations, not trajectories. We have updated the figure, which more accurately captures the density of observations and have added a 'rug plot' to the figure, with marks to indicate the observed WFL at months zero and 24.

It is still the case the trajectories tend to cluster into a smaller number of 'streams' as different initial conditions tend to converge, but to a much lesser extent.

**The discussion should expand on future research that the method leads to.**

**Anthropometry is a practical proxy for health and nutritional status and validation of criteria for intervention are typically based on mortality risk in large populations. As a next stage of research, prediction of actual health events such as death, serious illness or impaired neurodevelopment would be valuable to mention. Additionally, it would be helpful to comment on the models' ability to include time-varying covariates such as seasonality, illness or food security shock, as well as preterm birth.**

This is an interesting idea and perhaps the opposite way to how we were framing the question (looking at how events during childhood manifest in growth outcomes rather than growth as a predictor of later outcomes). Individual-level coefficients from the model could indeed be used as parameters to predict other health outcomes. We have noted the incorporation of seasonal parameters.

**The narrower CIs suggest the SDE model is a better approximation of the data than the linear mixed models. However, linear models may not be regarded as a current gold standard for growth models, and comparisons with others including latent growth models, SITAR, multilevel fractional polynomial models or penalised splines etc. This is a limitation of the paper and area for future work.**

It is not clear that there is a gold standard (linear mixed models are still widely used) as such and as Reviewer 1 notes, there is not much to pick between such models that are fundamentally similar in concept (e.g. fitting to the mean using some variant of least squares). We note that the accessibility of the method in conventional software might mean that this approach can be run in parallel for routine comparisons. That said, we have noted this important future avenue in the discussion.

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 04 August 2020

<https://doi.org/10.21956/gatesopenres.14305.r28999>

© 2020 Cameron N. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Noel Cameron**

School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough, UK

1. The authors need to explain why yet another model for growth is required or necessary or



advantageous.

2. Introduction: "One of the main barriers to analysing child growth data is that individual growth trajectories display highly variable and complicated dynamic behaviour, differing markedly between children, even from the same geographic and socio-economic group. As such, developing growth models from which actionable insights can be extracted – such as identification of interventions at the population level or predictive risk assessments at individual child level – is both methodologically and practically challenging."  
The variability demonstrated between children is not a "barrier" to analysing growth data. The variability is a measure of the variation in normal growth rates and in response to potential constraints on growth. To smooth out or diminish this variation without investigation destroys the very essence of growth studies particularly in clinical scenarios.
3. "WHO reference standards" – there is no such thing as a "reference standard". Growth charts are either based on cross-sectional large sample data and are "references" or selected longitudinal samples and are "standards". The WHO charts are standards.
4. Whilst I am impressed by the comparison between the OU method and LMM in Figure 4, the lack of any apparently significant improvements or insights into the growth trajectory makes me wonder if the new method really does represent a significant advantage over previous methods.
5. I would like the authors to explore what precisely is gained by applying this method as opposed to previous methods. Does it allow earlier detection of poor growth? Does it provide a more reliable prediction of future growth? If used as an intervention trigger what preceding data are required and prediction accuracy is obtained?
6. References include only minimal publications relating to growth modelling.

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

I cannot comment. A qualified statistician is required.

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Human growth and development - Auxology

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 17 Nov 2020

**Ben McCormick,**

**The authors need to explain why yet another model for growth is required or necessary or advantageous.**

This is an important consideration true of any modelling work. Given that models can serve a multitude of purposes in this case, we believe that this alternative modelling strategy (rather than the model results *per se*) is an opportunity to gain greater insight from anthropometric data. Growth models of the same data will (should!) arrive at very similar fitted values, however, the SDE approach is conceptually different (and has correspondingly different math). In this case, the SDE partitions variance between two different mechanistic phenomena and therefore offers an interpretation that is not available to more traditional curve-fitting approaches that describe the pattern, but perhaps not the underlying process. We have added text to the introduction to this effect.

**The variability demonstrated between children is not a “barrier” to analysing growth data. The variability is a measure of the variation in normal growth rates and in response to potential constraints on growth. To smooth out or diminish this variation without investigation destroys the very essence of growth studies particularly in clinical scenarios.**

We have revised the text to “challenges” and thank the reviewer for their comment. The principle purpose of an SDE model is that it explicitly describes variability in growth rates rather than fitting a population mean and treating variance as a nuisance parameter. SDEs are most often used where a process has noise and variability, as the reviewer notes, is the case with growth modelling. In this sense, the SDE is the perfect match to produce better (more robustly capturing uncertainty) forecasts of growth. We have amended the text to better draw out this point (“more accurately capturing external sources of variability and uncertainty”).

**“WHO reference standards” – there is no such thing as a “reference standard”. Growth charts are either based on cross-sectional large sample data and are “references” or selected longitudinal samples and are “standards”. The WHO charts are standards.**

Thank you for this clarification, we have amended the text accordingly.

**Whilst I am impressed by the comparison between the OU method and LMM in Figure 3, the lack of any apparently significant improvements or insights into the growth trajectory makes me wonder if the new method really does represent a significant advantage over previous methods.**

We illustrate an improved model fit using a relatively simple model akin to Brownian

motion. This manuscript lays out an introduction to the approach, but the SDE method is much richer in terms of capturing a dynamic stochastic process.

**I would like the authors to explore what precisely is gained by applying this method as opposed to previous methods. Does it allow earlier detection of poor growth? Does it provide a more reliable prediction of future growth? If used as an intervention trigger what preceding data are required and prediction accuracy is obtained?**

As with the first question posed by reviewer 2, we have noted that the principle advantage of the SDE approach is the descriptive partitioning of variance (“SDEs can enhance mechanistic interpretation of the drivers of variability (both long- and short-term) in growth”). In that sense, we believe this method can (with appropriate terms) offer superior mechanistic insight: variables can relate to one or both of the drift and shift components. With such insight, one might more appropriately distinguish between factors relating to growth trends and those relating to short-term fluctuations (albeit these fluctuations can have lasting consequences if negative effects aren’t removed).

**References include only minimal publications relating to growth modelling.**

Most published growth models take a very similar methodological approach (curve-fitting) whereas our intention was to introduce a conceptually different approach rather for contrast.

**Competing Interests:** No competing interests were disclosed.