ORIGINAL ARTICLE

Forecasting of cohort fertility under a hierarchical Bayesian approach

Joanne Ellison[1] | **Erengul Dodd**[1] | **Jonathan J. Forster**[2]

[1]University of Southampton, UK

[2]University of Warwick, UK

Correspondence
Joanne Ellison, Building 54 Mathematical Sciences, University of Southampton, Highfield, Southampton, SO17 1BJ, UK
Email: J.Ellison@soton.ac.uk

Funding information
The PhD programme of the first author is funded by the Engineering and Physical Sciences Research Council (award reference 1801045).

Fertility projections are a key determinant of population forecasts, which are widely used by government policymakers and planners. In keeping with the recent literature, we propose an intuitive and transparent hierarchical Bayesian model to forecast cohort fertility. Using Hamiltonian Monte Carlo methods and a dataset from the Human Fertility Database, we obtain fertility forecasts for 30 countries. We use scoring rules to quantitatively assess the predictive accuracy of the forecasts; these indicate that our model predicts with an accuracy comparable to that of the best-performing models in the current literature overall, with stronger performance for countries without a recent structural shift. Our findings support the position of hierarchical Bayesian modelling at the forefront of population forecasting methods.

KEYWORDS
Cohort fertility; Forecasting; Hamiltonian Monte Carlo; Hierarchical Bayesian models; Human Fertility Database; Scoring rules

1 | INTRODUCTION

Fertility is one of the three components of population change, together with mortality and migration - population forecasts are obtained by projecting these components forward under certain assumptions and methodologies. Government policymakers, decision makers and planners use these forecasts for numerous purposes such as planning for the future provision of basic societal needs, e.g., food, water and energy, as well as health and education services; shaping policies both locally and nationally; determining fiscal projections; and informing pensions models (Office for National Statistics, 2019a; Population Reference Bureau, 2001; National Records of Scotland, 2019). Fertility forecasts specifically are required to plan maternity and childcare services, predict demand for nursery school places, as

well as various other uses (Office for National Statistics, 2019b; Shang, 2012; National Records of Scotland, 2019). As a result, models that can generate plausible fertility forecasts with appropriate uncertainty are in high demand.

There is a large body of literature concerning the proposal of models to produce accurate estimates of fertility rates. Much of the early work focuses on parametric techniques, which involve choosing functions to fit closely to the bell-shaped curves of age-specific fertility rates (e.g., see Peristera and Kostaki (2007)). Such functions include polynomials (Brass, 1960), the Coale-Trussell function (Coale and Trussell, 1974), beta and gamma distributions (Hoem et al., 1981), and the Hadwiger distribution (Hadwiger, 1940). In addition, the relational models of Brass (1974) assume a linear relationship between an observed set of fertility rates and some standard (Booth, 2006). Hoem et al. (1981) describes a selection of these methods and gives a detailed comparison. Recently, a slight hump at younger ages has appeared in the curves for some developed countries (Chandola et al., 1999), which the previous models are unable to respond to (Peristera and Kostaki, 2007). Chandola et al. (1999) suggests a possible cause as the emerging differences between marital and non-marital fertility, proposing a mixture of Hadwiger functions to model this phenomenon.

Since the 1980s, attention has largely moved to models that treat fertility as stochastic rather than deterministic, and so are able to quantify the uncertainty in forecasts (Wiśniowski et al., 2015). The functional model of Lee (1992) is particularly notable in this school of thought, with its use of principal component analysis and time series inspiring many approaches that involve modelling the randomness of fertility (Booth, 2006). These further functional methods include the work of Hyndman and Ullah (2007) and Shang (2012). Consistent with the change to a probabilistic viewpoint, Bayesian models are now a popular choice in the population forecasting literature as they can incorporate uncertainty naturally. Recent papers such as Wiśniowski et al. (2015) and Bijak and Bryant (2016) attribute the rise in their usage to computational developments occurring as recently as the last decade. Hierarchical Bayesian models (e.g., see Girosi and King (2008)), which allow borrowing of strength, are also becoming increasingly common. This strength can be borrowed from other countries, an approach used in the methodology of the first probabilistic population projections to be published by the United Nations (Ševčíková et al., 2016). Alternatively, the strength can be borrowed across ages and cohorts (e.g., see Czado et al. (2005)) for the fertility rate estimates of a single country.

In an attempt to determine whether the increasingly sophisticated and computationally expensive models proposed in recent years have actually led to greater predictive accuracy, Bohk-Ewald et al. (2018a) perform a comprehensive comparison of 20 existing cohort fertility forecasting approaches. Taking a cohort approach, the aggregate fertility measure the authors use to compare the methods is the cohort total fertility rate (CFR). The CFR is calculated by summing the age-specific rates for a given cohort, i.e., a group of women with the same birth year, across all reproductive ages (Jasilioniene et al., 2015); as such, it can be interpreted as the average completed family size for that cohort. The equivalent measure under a period approach is the total fertility rate (TFR), which sums the rates for a given calendar year - this can be interpreted as the average completed family size for a *hypothetical cohort* of women who experience the fertility rates of that one year throughout their reproductive lives (Ní Bhrolcháin, 2011).

By its definition, the TFR provides a summary of fertility over a brief period of time which can be very recent, and is therefore more immediately relevant compared to the CFR (Bongaarts and Feeney, 1998). However, as well as the absence of a practical interpretation, a further drawback is that changes in the TFR can be due to tempo effects, i.e., changes in the average age of childbearing during the period in question, rather than just quantum effects, i.e., changes in the average number of children per woman (Bongaarts and Feeney, 1998). As a result, cohort fertility tends to be more stable across time than period fertility (de Beer, 1985; Li and Wu, 2003), which makes it a more appealing measure for forecasting purposes. For these reasons, combined with the frequent adoption of the cohort approach in the recent fertility forecasting literature, we also decide to take a cohort approach.

Returning to the work of Bohk-Ewald et al. (2018a), the authors find that in terms of forecast accuracy, four methods perform better than the naive freeze rates approach, which simply freezes the age-specific rates at their most recent observed values. These superior approaches include the two simple extrapolation methods of Myrskylä et al. (2013a) and de Beer (1985, 1989). The former extrapolates the age-specific trends exhibited over the previous five years for a further five years before freezing the rates; the latter extrapolates patterns exhibited by the rates across ages and cohorts jointly by fitting two interconnected ARIMA time series models. The remaining two successful approaches are both Bayesian methods, namely the conjugate normal-normal model of Schmertmann et al. (2014a) and the aforementioned model of Ševčíková et al. (2016). The former constructs a prior from quadratic penalties, simultaneously penalising potential future patterns of rates in the age and cohort dimensions that are deemed unlikely by the historical data; the latter first forecasts the TFR using a hierarchical Bayesian model, subsequently decomposing it into age-specific projections according to a particular target pattern determined by expert opinion.

Of the eight methods that allow uncertainty quantification (this includes the methods we have described except for that of de Beer (1985, 1989)), the Bayesian model of Schmertmann et al. (2014a) appears to perform strongest, and therefore could be seen as the best when considering forecast accuracy and uncertainty together. Overall, the questionable dominance of the Bayesian approaches over simple extrapolation methods causes the authors to question whether such complex models requiring large amounts of data and computation time are really necessary in order to obtain accurate cohort fertility forecasts - this is one of the motivations of our work.

This discussion brings us to the main purpose of this paper. In the spirit of the highly successful model of Schmertmann et al. (2014a) and in keeping with the most recent literature, we propose a hierarchical Bayesian model for forecasting cohort fertility. By incorporating our assumptions explicitly into the model structure and then letting the data determine their precise satisfaction, we aim to construct a transparent and intuitive model with realistic levels of forecast uncertainty that can compete with the current best-performing models in the field. We fit our model using the state-of-the-art Hamiltonian Monte Carlo computational methodology implemented by the software RStan (Stan Development Team, 2018a).

After presenting our approach in Section 2, in Section 3 we compare the forecast performance of our model with that of Schmertmann et al. (2014a), Myrskylä et al. (2013a) and de Beer (1985, 1989). We fit to the fertility data available in 2014 to generate forecasts for 30 countries and perform a qualitative comparison. We also fit to the data available 10 years earlier in 2004 to generate forecasts for 29 countries, allowing us to use scoring rules and other summary statistics to compare the approaches quantitatively. Lastly, we discuss our findings in Section 4.

2 | METHOD

2.1 | Introduction

Let us consider an incomplete Lexis surface, i.e., a heat map of fertility rate estimates plotted by age against cohort year of birth. For a cohort of women at a given age, the fertility rate is estimated by dividing the number of live births by the number of women (Jasilioniene et al., 2015). Figure 1 gives an example of a Lexis surface for England and Wales data from the Human Fertility Database (2019), taking the present to be 2014 by only using the data that would have been available in that year; also note that the surface is restricted to ages 15-44 and the 1923-1999

cohorts. Features of interest include the high rates in the dark region which occurred during the 1960s and early 1970s for women in their twenties, and the increase in the peak age of childbearing from the 1960s cohorts to the late 1970s cohorts (Office for National Statistics, 2017). We call the set of rates for each cohort a cohort schedule, with a cohort schedule complete if it is fully observed and incomplete otherwise. Using this terminology, we see from Figure 1 that the cohort schedules up to and including the 1970 cohort are complete - this is because in 2014, the 1970 cohort would have been the youngest cohort to have an observable rate for women aged 44.

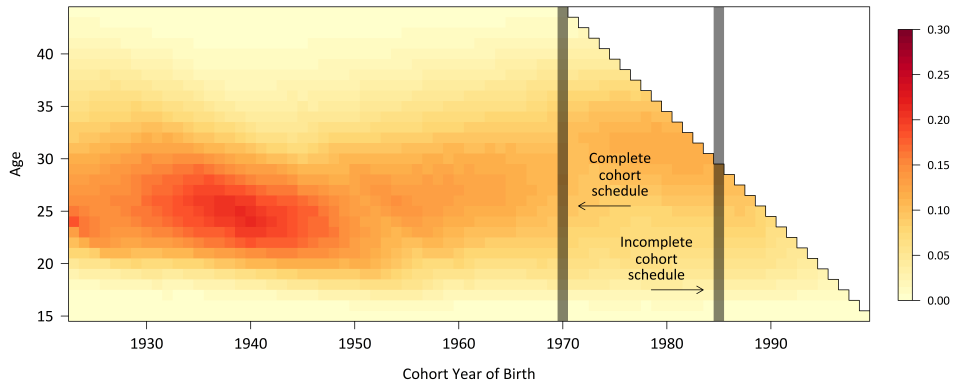

FIGURE 1 Lexis surface of England and Wales Human Fertility Database (2019) fertility rate estimates by age against cohort year of birth, taking the present to be 2014. White cells correspond to future rates which are yet to be observed. The last complete cohort schedule (of the 1970 cohort) is indicated, as well as an incomplete cohort schedule (of the 1985 cohort).

The dataset we consider as a whole consists of complete and incomplete cohort schedules for ages 15-44 from countries across Europe, with several countries from North America and Asia. Following Schmertmann et al. (2014a), we separate the data into historical and contemporary sections. For each country, we take all the available complete cohort schedules for the 1904-1953 cohorts to form the historical part, and the 1960-1999 cohort schedules to form the contemporary part. We illustrate this in Figure 2 for England and Wales. From Figure 2a, we see that England and Wales only contributes 31 of the 50 desired historical cohort schedules from the 1923 cohort onwards.

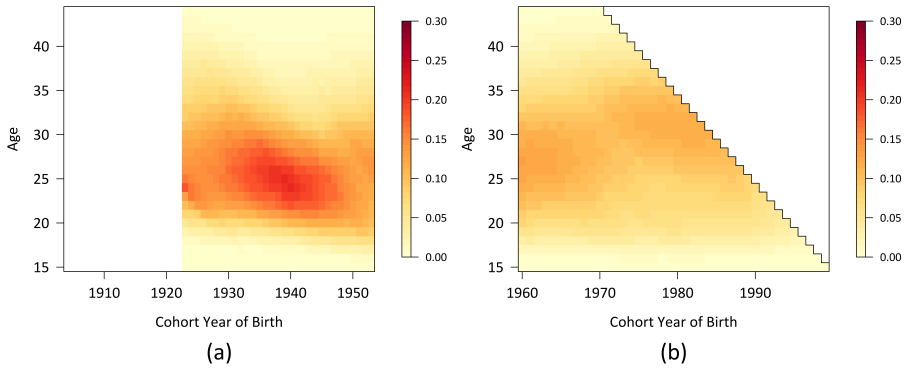

FIGURE 2 England and Wales (a) historical and (b) contemporary Lexis surfaces (Human Fertility Database, 2019).

The reason for this separation is apparent upon considering the model structure. For a given country, the parameters are the true fertility rates in its contemporary Lexis surface (one for each of the possible cohort-age combinations in Figure 2b). The historical data from all the countries (displayed in Figure 3) informs the core of the model. We specify a prior for the (hyper)parameters, which is then updated by combining its information with that obtained from the country's observed contemporary rates (the rate estimates in Figure 2b) in the likelihood. This gives a posterior distribution for the parameters of the entire contemporary Lexis surface. We specify our model in Section 2.2.

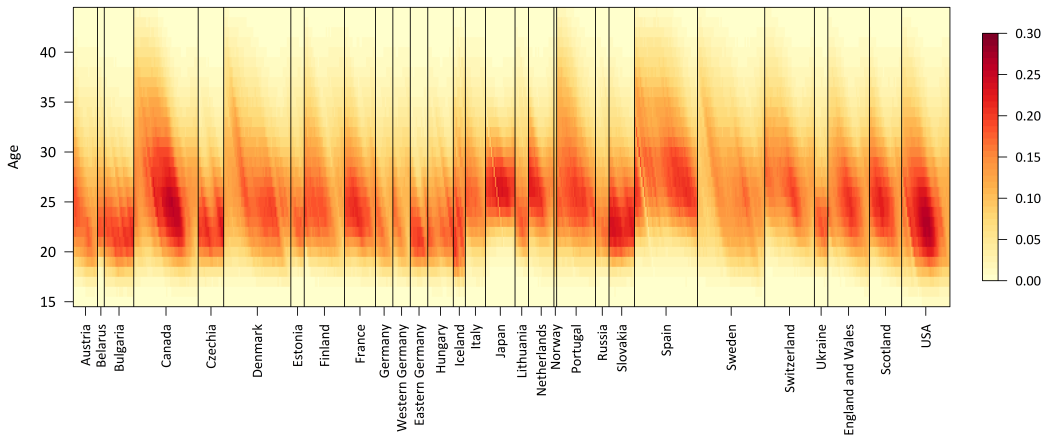

FIGURE 3 The combined historical Lexis surfaces from all of the countries in our dataset (Human Fertility Database, 2019).

2.2 | Model specification

Suppose there are C birth cohorts ($c = 1, \ldots, C$) and A ages ($a = 1, \ldots, A$) in the contemporary Lexis surface of a particular country. For each cohort-age combination (c, a), let N_{ca} be the observed number of births, W_{ca} be the number of women alive (i.e., the exposure) and θ_{ca} be the true fertility rate; then $\mu_{ca} = W_{ca}\theta_{ca}$ is the mean number of births. Due to its suitability for modelling count data, we assume a Poisson distribution for N_{ca} with mean μ_{ca}, i.e., $N_{ca} \sim \text{Poisson}(\mu_{ca})$. We model θ_{ca} on the logarithmic scale, the standard approach to take when modelling rates. It also ensures that our model will not generate negative forecasts, which is especially beneficial in instances where we have diminishing fertility at ages where the level is already low. Letting H be the total number of complete historical cohort schedules contributed by all the countries, we define $\mathbf{\Phi}$ to be the $A \times H$ matrix of these cohort schedules (see Figure 3 in Section 2.1 for a visual representation of such a matrix with $A = 30$ and $H = 653$). Let $\mathbf{\Pi}$ be the equivalent matrix on the logarithmic scale, i.e., $\mathbf{\Pi}_{ah} = \log(\mathbf{\Phi}_{ah}), a = 1, \ldots, A, h = 1, \ldots, H$. We then let $\mathbf{X}$ be the $A \times 3$ matrix consisting of the first three principal components of $\mathbf{\Pi}$, obtained from its singular value decomposition (SVD). These components are essentially highly significant covariates that together explain a large proportion of the variation of the historical cohort schedules. The underlying assumption, also made in Schmertmann et al. (2014a), is that these components will explain a similar proportion of the variation in the contemporary cohort schedules. We now define the core of the model, which takes the following log-linear form:

$$\log(\theta_{ca}) = [\mathbf{X}\boldsymbol{\beta}]_a + \varepsilon_{ca}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ is the vector of regression parameters, i.e., the respective weights on the three principal components - see Schmertmann et al. (2014a) for an interpretation of these weights when the principal components of $\boldsymbol{\Phi}$ are computed. The β_i's are allowed to vary freely, imposing no constraints on the possible shapes or levels of the incomplete cohort schedules; we ensure that this is the case by giving each β_i a diffuse $N(0, 30^2)$ prior. The ε_{ca}'s are the error terms, and we enforce our remaining model assumptions through their prior distribution.

Firstly, letting $\boldsymbol{b} = (b_1, b_2, b_3)'$, from (1) it is clear that the model is invariant under the following parameterisation:

$$\{\boldsymbol{\beta}, \varepsilon_{ca}\} \rightarrow \{\boldsymbol{\beta} + \boldsymbol{b}, \varepsilon_{ca} - [\mathbf{X}\boldsymbol{b}]_a\}.$$

To resolve this identification problem, we impose a constraint on the vector of 'jump-off' ε_{ca}'s, i.e., the ε_{ca}'s whose cohort-age combinations correspond to the calendar year we are taking to be the present - we will discuss the reason for this choice in due course. We call this vector ε_{JO} and illustrate it in Figure 4a below. We let $\varepsilon_{JO} = (\mathbf{I}_A - \mathbf{P}_X)\boldsymbol{\eta}$, where $\boldsymbol{\eta} = (\eta_1, ..., \eta_A)'$ and $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the projection matrix of $\mathbf{X}$. We then let $\eta_a \overset{\text{iid}}{\sim} N(0, \sigma_1^2)$, $a = 1, \ldots, A$. Hence $\mathbf{X}'\varepsilon_{JO} = 0$. As $\mathbf{X}$ has three columns, this imposes three linear constraints on ε_{JO}, fixing its level and therefore making the model identifiable.

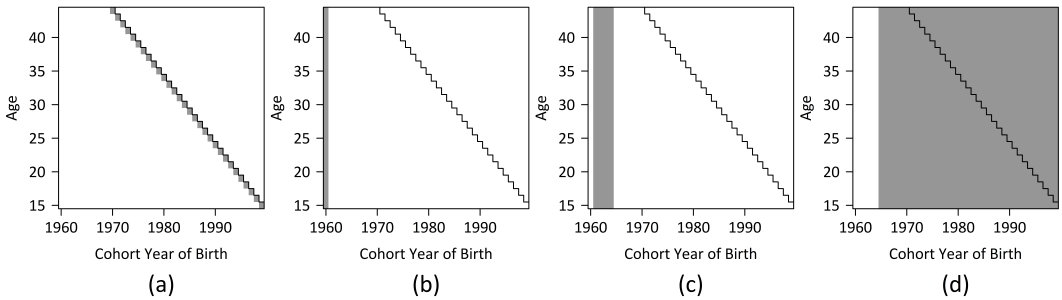

FIGURE 4 Representation of (a) ε_{JO}, (b) Region A, (c) Region B and (d) Region C on a typical contemporary Lexis surface where the present is 2014, with the included cohort-age combinations filled in grey.

Next, we divide the contemporary Lexis surface into three regions A-C, which are illustrated in Figures 4b-d. Region A consists of the ε_{ca}'s with cohort-age combinations in the first cohort only, Region B those in cohorts two to five, and Region C those in cohort six onwards. We then specify the priors on the ε_{ca}'s by the region they belong to:

Region A : Let $\varepsilon_{1a} \overset{\text{iid}}{\sim} N(0, \sigma_2^2)$, $a = 1, \ldots, A$

Region B : Let $\varepsilon_{ca} \sim N(\rho_1 \varepsilon_{(c-1)a}, \sigma_3^2)$, $c = 2, \ldots, 5$

Region C : Let $\varepsilon_{ca} \sim N(\rho_2 \varepsilon_{(c-1)a} + \rho_3(\varepsilon_{(c-1)a} + \hat{\Delta}_{(c-1)a}), \sigma_4^2)$, $c = 6, \ldots, C$,

where $\hat{\Delta}_{(c-1)a} = \frac{1}{30}(10\varepsilon_{(c-1)a} - \varepsilon_{(c-2)a} - 2\varepsilon_{(c-3)a} - 3\varepsilon_{(c-4)a} - 4\varepsilon_{(c-5)a})$ is the ordinary least squares (OLS) slope estimator obtained by fitting a linear regression model without an intercept to the five error terms corresponding to age a and cohorts $c-5, \ldots, c-1$.

Lastly, we assume that each $\sigma_i \sim N^+(0, 0.25^2)$ and each $\rho_i \sim N^+(0, 0.5^2)$, where $N^+(\cdot, \cdot)$ indicates a half-normal prior.

In words, our Region A prior states that the error terms are simply independently and identically normally distributed with mean 0 and variance σ_2^2. Our Region B prior states that for each age, the ε_{ca}'s follow an AR(1) process across cohort with coefficient ρ_1 and constant error variance σ_3^2. In Region C, the prior states that the ε_{ca}'s are normally distributed with mean equal to a weighted combination of the previous age-specific error $\varepsilon_{(c-1)a}$, and the sum of this error term and the slope $\hat{\Delta}_{(c-1)a}$ that it gives rise to along with the previous four age-specific errors, with weights ρ_2 and ρ_3 respectively and variance σ_4^2. The Region C prior is the most important as it will determine the forecasts - this is because this region includes all the future cohort-age combinations. What we are actually doing in this prior is balancing the two most common extrapolation methods for observed fertility rates in the demographic forecasting literature, which we will call the freeze-rate and freeze-slope approaches respectively in line with Schmertmann et al. (2014a). The freeze-rate approach assumes that the next age-specific rate will be similar to the previous one, i.e., $\theta_{ca} \approx \theta_{(c-1)a}$. On the other hand, the freeze-slope approach assumes that the next age-specific rate will follow the recent trend of its past rates, i.e., $\theta_{ca} \approx \theta_{(c-1)a} + \hat{\delta}_{(c-1)a}$, where $\hat{\delta}_{(c-1)a}$ is the recent slope - we take this to be calculated using the last five rates in the same spirit as Myrskylä et al. (2013a) and Schmertmann et al. (2014a).

In our model we are working on a logarithmic scale and with these assumptions applied to the error terms instead of the actual rates; these are two of the key differences from the model of Schmertmann et al. (2014a). For the freeze-rate approach the two are equivalent using (1), i.e., $\varepsilon_{ca} \approx \varepsilon_{(c-1)a} \implies \theta_{ca} \approx \theta_{(c-1)a}$. However for the freeze-slope approach they are not, as $\varepsilon_{ca} \approx \varepsilon_{(c-1)a} + \hat{\Delta}_{(c-1)a} \implies \theta_{ca} \approx \theta_{(c-1)a} \times \exp(\hat{\Delta}_{(c-1)a})$, again using (1). This is intuitive, as a small change on the log scale is approximately equivalent to the proportionate change on the original scale.

Returning to the Region C prior, it is now clear that we are allowing the data to choose how much weight to put on the freeze-rate and freeze-slope assumptions through ρ_2 and ρ_3 respectively. We do not constrain these parameters to sum to 1, as if $\rho_2 + \rho_3 < 1$ each age-specific process is stationary and will revert to $[\mathbf{X}\boldsymbol{\beta}]_a$ in the long term. If $\rho_2 + \rho_3 > 1$ then the process is non-stationary and will not exhibit long-term reversion, instead having a more explosive nature. By leaving the sum of ρ_2 and ρ_3 unconstrained, for each country we allow the parameters to learn from the observed contemporary fertility rate estimates, choosing whether they want to follow a stationary or non-stationary process as a result. The degree of stationarity, i.e., how close $\rho_2 + \rho_3$ is to 1, is also significant, as it determines how strongly the reversion occurs in the forecasts. It is this reversion of a stationary process to $[\mathbf{X}\boldsymbol{\beta}]_a$ that motivated our decision to constrain ε_{JO} to make the model identifiable earlier in this section. The rationale behind this is that if we could choose where we would want our forecasts to revert to in the future, it would be somewhere close to where we started forecasting from, i.e., the value in the calendar year taken to be the present, as opposed to the initial value or some average across the contemporary cohort schedules. This is in line with the current fertility literature, e.g., the use of only the last five years of data in Myrskylä et al. (2013a). In constraining ε_{JO} therefore, the desire is that for each country $[\mathbf{X}\boldsymbol{\beta}]_a$ will be close to the jump-off value for each age a. We discuss this further in Section 3.2.3.

Here we propose a hierarchical model to borrow strength across ages and cohorts. This is particularly important in Region C, where we allow ρ_2, ρ_3 and σ_4^2 to learn from all the observed cohort-age combinations after the fifth cohort. In this way, our forecasts take as much information as possible from the contemporary Lexis surface about the relative importance of the freeze-rate and freeze-slope assumptions. Our model is also hierarchical in the typical sense, in that we have two levels of priors due to the presence of the hyperparameters (the σ_i's and the ρ_i's). We give an overview of the model fitting in Section 2.3.

2.3 | Model fitting

The hierarchical nature of our proposed model means that it is not possible to write the posterior in closed form - instead, we need to approximate it using Monte Carlo methods. Such methods are also required for the hierarchical Bayesian model of Ševčíková et al. (2016), whereas the posterior of the conjugate Bayesian model of Schmertmann et al. (2014a) is tractable and hence can be computed precisely. For complex hierarchical models such as ours, well-known Markov chain Monte Carlo (MCMC) methods like the Metropolis algorithm are less satisfactory due to their local random walk behaviour, i.e., slow exploration of the posterior (e.g., see Gelman et al. (2014)). The method of Hamiltonian Monte Carlo (HMC) increases the efficiency of this exploration through Hamiltonian dynamics (for more details see Stan Development Team (2018b)). The variant of HMC that we will use in the computation of the proposed model is the no-U-turn sampler (NUTS), which determines certain algorithm parameters adaptively in each iteration so as to maximise the exploration distance relative to the current position. The NUTS is implemented by the software RStan (Stan Development Team (2018a)), which we will use to fit our model in Section 3.

The fitting process consists of two parts. First, we use RStan to perform T iterations of the NUTS algorithm, following a warmup period of T' iterations for the purposes of estimation and optimisation of algorithm parameters. This generates T samples of β, the σ_i's and the ρ_i's, as well as the ε_{ca}'s with observed cohort-age combinations and therefore observed values of N_{ca} and W_{ca}. Second, for the purposes of forecasting we need to obtain T samples of the ε_{ca}'s with unobserved cohort-age combinations. We do this by simulating them from $N(\rho_2\varepsilon_{(c-1)a} + \rho_3(\varepsilon_{(c-1)a} + \hat{\Delta}_{(c-1)a}), \sigma_4^2)$, one set of samples for each of the T original samples. This gives T samples of $\theta_{ca} = \exp([\mathbf{X}\boldsymbol{\beta}]_a + \varepsilon_{ca})$ for each cohort-age combination. We are able to code our model in such a way that RStan can perform this simulation within each iteration.

The posterior distribution allows us to quantify our uncertainty about the true rates θ_{ca} but not the empirical birth rates. We need to incorporate the additional variation to account for the fact that we are predicting an observation and not the mean. The process by which we do this for a conjugate model (e.g., Schmertmann et al. (2014a)) is described in Appendix A, and is simple because we have the posterior distribution in closed form. The process is slightly more involved for the proposed model however, as we only have T samples of each θ_{ca} available to us. Our goal is to have a credible interval for each empirical rate as these are what we are trying to forecast; for this reason we need to generate empirical birth rates from our T samples of each true birth rate θ_{ca}. For each (c, a), we do this by first sampling a random observation from Poisson(μ_{ca}^t) for $t = 1, ..., T$, where $\mu_{ca}^t = W_{ca}\theta_{ca}^t$ is the tth sampled value of μ_{ca} and θ_{ca}^t the tth sampled value of θ_{ca}. If (c, a) is unobserved, we take W_{ca} to be its most recently observed value at age a. We then divide each of these T Poisson realisations by W_{ca} to obtain a sample of T empirical birth rates. We then compute the 90% and 50% credible intervals (CIs), which are probability intervals based on the posterior predictive distribution. We do this by extracting the $(5\%, 95\%)$ and $(25\%, 75\%)$ quantiles of the sample of empirical birth rates. Note that the additional uncertainty only has a noticeable effect for small countries with comparatively low exposures.

3 | RESULTS

3.1 | Data and computation

In order to assess the forecast performance of our proposed model, we follow the advice of Bohk-Ewald et al. (2018a) to compare against the naive freeze rates method and the simple extrapolation models of Myrskylä et al. (2013a) and de Beer (1985, 1989) at a minimum; we will refer to these as Models MGC and dB respectively, after the authors. We additionally include the model of Schmertmann et al. (2014a) in our comparison (denoted Model SZGM), as our proposed hierarchical Bayesian model (denoted Model hB) has been developed in the same spirit. We fit the models to the countries available in the Human Fertility Database (2019)[1] dataset. We first fit the models to the data available in 2004, allowing us to use the more recent (or holdout) data to perform a quantitative comparison using scoring rules and various summary statistics relating to point *and* probabilistic accuracy. This generates forecasts for 29 countries, which we call '2004 (fertility) forecasts' and present in Section 3.2. We then incorporate the holdout data directly by fitting the models to the data available in 2014, generating '2014 (fertility) forecasts' for 30 countries which we discuss in Section 3.3. The R code and Stan files to obtain the Model hB forecasts are available from `https://github.com/jvellison/hBfert`. For the Model SZGM computation we use the R code available on the Schmertmann et al. (2014a) project website (Schmertmann et al., 2014b), slightly modified to account for the change in data and incorporation of additional variation (see Appendix A). For Model MGC we use our own R code written according to the method described in Myrskylä et al. (2013a) and inspired by the corresponding Stata code (Myrskylä et al., 2013b) as well as the R code used to produce the work of Bohk-Ewald et al. (2018a) (Bohk-Ewald et al., 2018b); we also use this R code to implement Model dB. We fit Model hB to each country separately, with $T' = 1,000$ warmup iterations followed by $T = 4,000$ retained iterations (see Section 2.3) thinned by a factor of two from an initial $8,000$. We examine convergence for each fit in the conventional way, and find that the samples mix well across the T iterations.

3.2 | 2004 fertility forecasts

In this section we provide an analysis of the 2004 forecasts using various approaches. First, in Section 3.2.1 we use scoring rules to quantitatively compare the accuracy of the Model hB, SZGM, MGC, dB and freeze rates forecast distributions. In Section 3.2.2 we use typical summary statistics to compare their point accuracy and coverage (where possible). Then, in Section 3.2.3 we graphically explore both the stationarity of the Model hB forecasts and the way in which the degree of stationarity affects the nature of the reversion.

3.2.1 | Scoring rules

To compare the models on their forecast precision and uncertainty we use scoring rules, which are measures of predictive accuracy for probabilistic prediction (Gelman et al., 2014). This means that they consider the posterior distribution as a whole rather than a summary statistic of it such as the mean or median. For a probabilistic forecast G (with associated CDF G and PDF g) and observed value y, a scoring rule summarises the suitability of G in light of y by a

[1] Note that this is an update to the 2011 version of the dataset used in Schmertmann et al. (2014a). It includes 12 additional countries (Belarus, Chile, Croatia, Denmark, Iceland, Italy, Japan, Norway, Poland, Spain, Taiwan and Ukraine), modifications to rate estimates available in 2011, and additional rate estimates.

score. The scoring rules we will use are negatively oriented, which means that smaller scores are desirable (Jordan et al., 2017). Following Jordan et al. (2017) we compute the logarithmic score (LogS; Good (1952)) and the continuous ranked probability score (CRPS; Matheson and Winkler (1976)):

$$\text{LogS}(G, y) = -\log(g(y)); \tag{2}$$

$$\text{CRPS}(G, y) = \int_{\mathbb{R}} (G(z) - \mathbb{I}\{y \leq z\})^2 \text{dz}, \tag{3}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. From equation (2), it is clear that the LogS penalises forecast distributions that assign negligible probability to y. In contrast, the CRPS "generalises the absolute error" (Gneiting and Raftery, 2007) by penalising forecast distributions whose CDFs differ substantially from the empirical CDF of y, i.e., the "perfect forecast" (Bröcker, 2012). In fact, equation (3) reduces to the absolute error if G is a deterministic (or point) forecast; therefore the CRPS allows us to score probabilistic and deterministic forecasts under the same metric, which is an attractive property (Gneiting and Raftery, 2007). Consequently, it is not surprising that the CRPS is more "sensitive to distance" (Gneiting and Raftery, 2007) and so would score a narrow distribution with median close to y, but negligible probability assigned to y, more favourably than the LogS - this is because the forecast *is* accurate, even though it is too precise. For further discussion of scoring rules, see Gneiting and Raftery (2007).

Regarding computation for the models with probabilistic forecasts (Models hB, SZGM and MGC), the LogS and CRPS can be calculated exactly under Models SZGM and MGC due to their Gaussian forecast distributions; the same is not true for Model hB due to its intractable posterior (see Section 2.3), and as a result approximations are required. We use a Gaussian approximation for g to compute the LogS and an empirical CDF-based approximation for G to compute the CRPS (see Krüger et al. (2019) for details).

We fit the models over ages 15-44, using the 1950-1989 cohorts as our contemporary data and maintaining the 1904-1953 cohorts as our historical data as in Section 2.1. We fix the number of complete contemporary cohorts at 11 (meaning that the 1950-1960 cohorts are complete and the 1961-1989 cohorts are increasingly incomplete), and focus solely on the CFR for simplicity and consistency with the recent literature. For each model we compute the CRPS for the CFR forecasts with a corresponding observed value - the number of such forecasts varies by country due to data availability, ranging from 5-13 with a modal value of 12 observed CFR values available after 2004 (for the 1961-1972 cohorts). We then plot the average CRPS by country in Figure 5, noting that this reduces to the mean absolute error for freeze rates and Model dB as their forecasts are deterministic. We order the countries by decreasing average CRPS (increasing predictive accuracy) under Model hB, for ease of comparison against the other methods.

The figure provides strong support for Model hB being competitive with the current best cohort fertility forecasting methods - its average score is significantly better than that for freeze rates for 27 of the 29 countries, and only marginally worse for the remaining two countries (Iceland and Bulgaria). In terms of its performance among the models identified as the most accurate in Bohk-Ewald et al. (2018a), it is fair to say from inspecting Figure 5 that for all countries from Denmark to the right (excluding England and Wales and Slovakia), the difference between the Model hB average score and the lowest for that particular country (achieved by Model hB for Portugal and the USA) is negligible. For the countries left of Denmark we see larger differences from the minimum, in particular for Switzerland, Iceland and Estonia; Model SZGM exhibits the opposite trend, performing very strongly for these countries overall but poorly for Denmark, Sweden, Finland, the USA, the Netherlands and Spain. To facilitate a fair comparison, we present histograms of the average scores for each model in Appendix B (Figure 12) - the Model hB and SZGM plots indicate that, excluding the Netherlands, their average CRPS distributions across countries are quite similar.

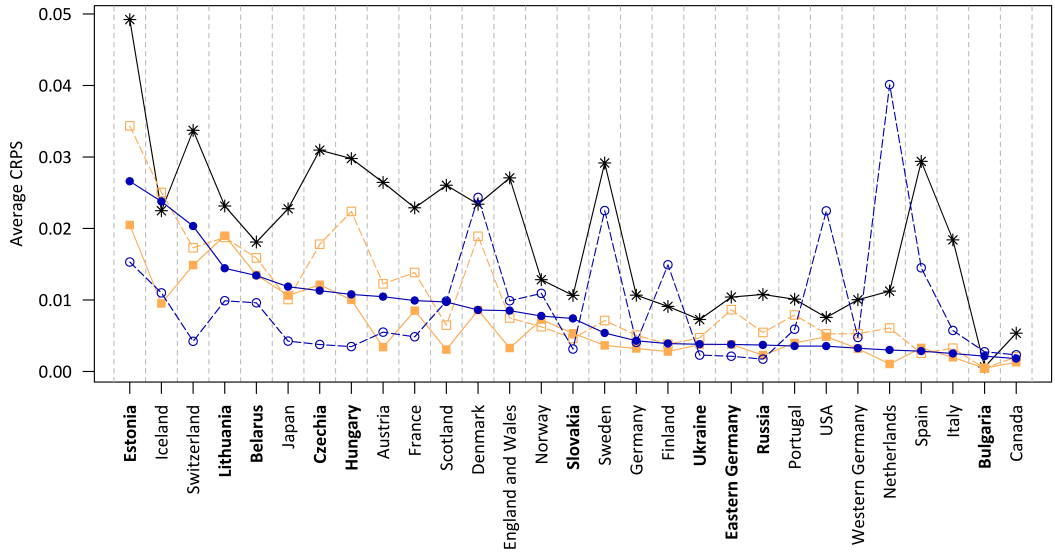

FIGURE 5 Plot of the average continuous ranked probability score (CRPS) against country for the 2004 cohort total fertility rate (CFR) forecasts for the proposed model (hB), the models of Schmertmann et al. (2014a) (SZGM), Myrskylä et al. (2013a) (MGC) and de Beer (1985, 1989) (dB), and the naive freeze rates approach. Model hB—●; Model SZGM—○; Model MGC—■; Model dB—□; freeze rates—✳. Countries are ordered by decreasing average CRPS under Model hB; post-communist countries are in **bold**.

Of the two simple extrapolation models, Model MGC clearly outperforms Model dB as well as the two Bayesian models, with a highly competitive average score across nearly all the countries; this conclusion is also supported by the relevant histograms in Figure 12. This strong performance is consistent with the assessment of Bohk-Ewald et al. (2018a) in terms of forecast accuracy, however the authors did find its forecast uncertainty to be substantially weaker in comparison. From the discussion of scoring rules at the start of this section, we know that the CRPS scores accurate forecast distributions with poor coverage more favourably compared to the LogS, which explicitly penalises forecasts which assign small probabilities to the true value. It is therefore also important to consider how Models hB, SZGM and MGC perform under the LogS, which we do by presenting Figure 6 (the equivalent of Figure 5 for the LogS) overleaf.

The most notable difference between the trends exhibited in Figures 5 and 6 is the highly erratic and unpredictable nature of the Model MGC scores in the latter - indeed, we are unable to display some of the Model MGC average LogS values as they are too large. This dramatically poorer performance for Model MGC under the LogS compared to the CRPS indicates that its forecast distributions grossly underperform in terms of forecast uncertainty, agreeing with Bohk-Ewald et al. (2018a). Regarding the two Bayesian models, again we see evidence of their complementary behaviour in that where Model SZGM performs badly, Model hB tends to perform well and vice versa. Under this scoring rule the models appear to be more balanced in terms of the relative magnitudes of their differences from the smallest average score, excluding the poor performance for the Netherlands under Model SZGM; this improvement for Model SZGM suggests that overall, Model hB may perform slightly worse in terms of coverage but better in terms of forecast accuracy. In Section 3.2.2 we will investigate whether the summary statistics support these conclusions.

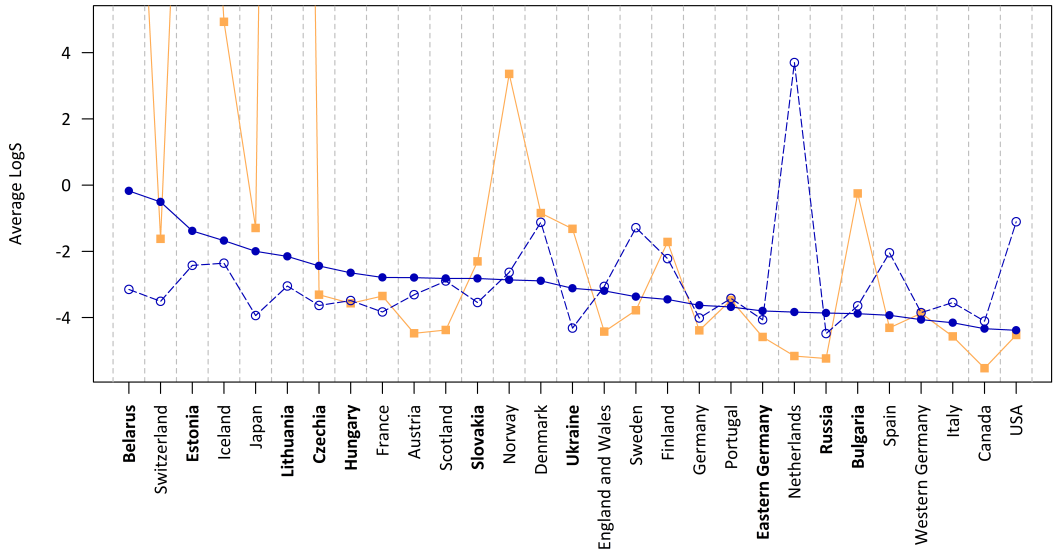

FIGURE 6 Plot of the average logarithmic score (LogS) against country for the 2004 cohort total fertility rate (CFR) forecasts for the proposed model (hB), the models of Schmertmann et al. (2014a) (SZGM) and Myrskylä et al. (2013a) (MGC). Model hB—●; Model SZGM—○; Model MGC—■. Countries are ordered by decreasing average LogS under Model hB; post-communist countries are in **bold**.

To get some idea of what the forecast distributions actually look like, we present the Model hB and SZGM forecast distributions graphically for five countries in Figure 7; note that we do not present the Model MGC forecast distributions in these plots for simplicity and due to their pre-established poorer coverage. The first two plots (from L-R) of Figure 7 represent countries that perform significantly better under Model hB compared to Model SZGM according to the scoring rules (the Netherlands and the USA). This is evident from the way that the holdout CFR values fall in the centre of the Model hB 50% CIs, whereas they drift from the Model SZGM intervals after the first few forecast years. The next two plots are where the opposite is true, i.e., Model SZGM outperforms Model hB in the scoring rules. The

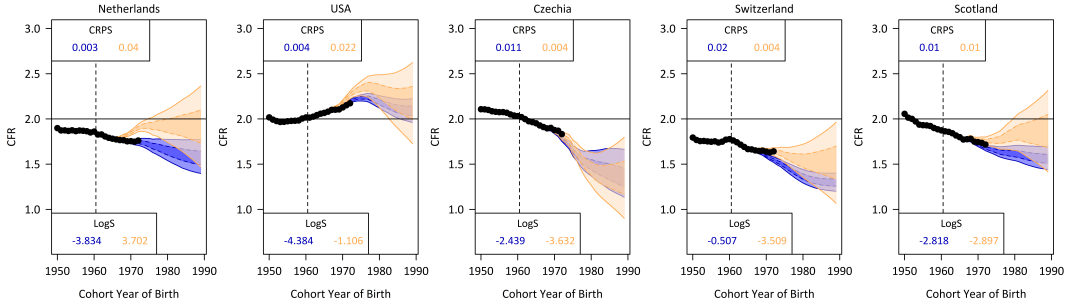

FIGURE 7 2004 cohort total fertility rate (CFR) posterior distributions for selected countries, with average continuous ranked probability score (CRPS) and logarithmic score (LogS) for the proposed model (Model hB) and the model of Schmertmann et al. (2014a) (Model SZGM) respectively and the dashed line indicating the start of the forecast period: Model hB 90% credible interval (CI)—■; Model hB 50% CI—■; Model SZGM 90% CI—■; Model SZGM 50% CI—■; Human Fertility Database (2019)—●.

first of these, Czechia, is a post-communist (PC) country, and the effect of the declining rates experienced across the post-communist region (Billingsley, 2010) on the forecast CIs is clear. Both models seemingly project the downturn unrealistically into the future, with the wider CIs for Model SZGM being the sole reason why it outperforms Model hB. The second case for Switzerland is more convincingly in favour of Model SZGM, looking like the reverse of the first two plots. Lastly, the right plot for Scotland gives an instance where there is little to choose between the models under both scoring rules; the CIs overlap for the first few forecast years before diverging, with the subsequent holdout CFR values falling roughly in between them. We also observe that the Model hB credible intervals are consistently narrower than those for Model SZGM - we will see whether this difference leads to a reduction in coverage compared to Model SZGM when we examine the summary statistics in Section 3.2.2.

Overall, through the use of scoring rules to assess the predictive performance of the Model hB cohort fertility forecast distributions relative to some of the current best models in the field, we have found strong evidence to support Model hB being competitive in terms of forecast accuracy *and* uncertainty. Figure 13 in the on-line supporting information gives the CFR plots for the countries not represented in Figure 7.

3.2.2 | Summary statistics

The scoring rules in Section 3.2.1 are single metrics that are able to quantify the overall performance of a forecast distribution in terms of accuracy and uncertainty; next we use various standard summary statistics in order to assess these two qualities separately, with the results presented in Table 1. Note that for simplicity and ease of interpretation, these statistics are calculated across all the countries rather than being country-specific as the average scores were in Section 3.2.1 - in this way we can determine whether these results are consistent with our previous general findings.

TABLE 1 Summary statistics calculated across all countries for the 2004 cohort total fertility rate forecasts under the proposed model (hB), the models of Schmertmann et al. (2014a) (SZGM), Myrskylä et al. (2013a) (MGC) and de Beer (1985) (dB), and the naive freeze rates approach; MAE = mean absolute error; MAPE = mean absolute percentage error; RMSE = root mean square error; CI = credible interval.

Measure	hB	SZGM	MGC	dB	Freeze rates
MAE (3dp)	0.011	0.013	0.009	0.011	0.020
MAPE (%, 2dp)	0.63	0.72	0.49	0.62	1.15
RMSE (2sf)	0.021	0.024	0.016	0.021	0.034
Coverage of 90% CI (%)	76	83	56	—	—
Coverage of 50% CI (%)	58	54	32	—	—

We have given three measures of predictive accuracy, namely the mean absolute error (MAE), mean absolute percentage error (MAPE) and the root mean square error (RMSE); note that for the models with probabilistic forecasts, we compute each error using the median of the relevant forecast distribution. The MAE values for the four models (excluding freeze rates due to its MAE being substantially larger) indicate that the typical magnitude of the CFR forecast error is 0.01, i.e., 10 children for every 1000 women in a given cohort over their reproductive lives. Freeze rates actually performs worst by a long way under all three measures, which is not surprising given the analysis of Figure 5 in Section 3.2.1 and the findings of Bohk-Ewald et al. (2018a). Focusing on the four models, we see that Model MGC

has the lowest value (and therefore best forecast accuracy) for all three statistics; this confirms our conclusions based on the computation of the CRPS in Section 3.2.1. In terms of the ordering of the remaining models, Models hB and dB have nearly identical values, all slightly larger than those for Model MGC; model SZGM is a little further behind again. This equivalence of Models hB and dB does not necessarily follow from Section 3.2.1, where Model dB appeared to perform significantly worse than Model hB from Figure 5 - however, the fact that the scoring rules are not a measure of forecast accuracy alone (they also take into account forecast uncertainty) could explain this difference.

The reason for Model SZGM performing relatively badly under these measures is likely due to the frequent high scores in Figure 5 pulling up the averages. When we consider the results for just the non-post-communist (non-PC) countries (not presented in Table 1), they actually show a greater margin of improvement for Model hB over Model SZGM across the three statistics - in particular, the MAPE decreases to 0.56% for Model hB while it increases to 0.85% for Model SZGM). Naturally this is countered by the opposite effect being observed for the PC countries (here the MAPE increases to 0.78% for Model hB while it decreases to 0.45% for Model SZGM, the lowest across all the models). So there appears to be evidence that Model hB is better suited to forecasting the CFR for countries with more stable contemporary fertility histories.

To quantify uncertainty we compute the coverage of the 90% and 50% CIs for the models with probabilistic forecasts (i.e., Models hB, SZGM and MGC). The results are consistent with our findings from computing the logarithmic score in Section 3.2.1, with Model SZGM closest to the nominal values, followed closely by Model hB and then Model MGC, which has very poor coverage. In terms of the coverages for the non-PC and PC countries separately, we see a similar trend to that observed for the predictive accuracy measures - the Model hB coverages exceed those for Model SZGM for the non-PC countries (79% versus 77% and 61% versus 44%), but are substantially lower for the PC countries (71% versus 96% and 54% versus 75%). This provides further evidence for our previous conclusion that the forecast performance of Model hB is more favourable for the countries without a recent structural shift.

To summarise, the analysis in this section has confirmed our findings from Section 3.2.1, that Model hB has forecast accuracy comparable to that of the current best cohort fertility forecasting models - indeed, only Model MGC performs significantly better. Conclusions are harder to state with forecast uncertainty, as we only have the strong and weak coverage of Models SZGM and MGC respectively (established in Bohk-Ewald et al. (2018a)) to compare against. Model hB lies in between the two models in this regard, but is undoubtedly closer to Model SZGM than Model MGC; it is most competitive with Model SZGM when considering the non-PC countries alone, where its coverage is slightly higher than that of Model SZGM. Overall these results are positive for Model hB, however the poorer performance for PC countries is concerning and should be investigated - we do this in Section 3.2.3.

3.2.3 | Stationarity and reversion

Lastly, we return to the stationarity discussion in Section 2.2 by presenting the posterior distributions of ρ_2, ρ_3 and their sum by country in Figure 8. First, we note that for each country, the ρ_2 error bars lie beneath the ρ_3 error bars (comparing Figures 8a and 8b). This means that the observed time series of age-specific rates in the contemporary Lexis surfaces are telling us that more weight should be put on the freeze-slope approach, i.e., following the recent age-specific trends, compared to the freeze-rate approach, i.e., remaining at the current age-specific level. Another interesting point is that all the countries choose for their age-specific processes to be stationary, which we can see from the $\rho_2 + \rho_3$ error bars in Figure 8c all lying below 1. This means that our age-specific forecasts all revert to $[\mathbf{X}\beta]_a$

in the long term and so are unlikely to be explosive. Despite this, there does appear to be a difference between the average degree of stationarity, i.e., how close the sum is to 1, for the PC and non-PC countries. The former tend to have their distributions of $\rho_2 + \rho_3$ at a lower level, and hence exhibit a faster reversion.

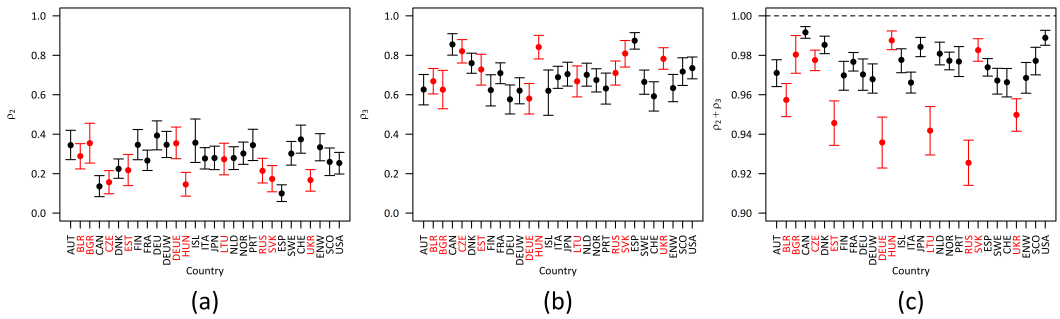

FIGURE 8 2004 posterior distribution summary of (a) ρ_2, (b) ρ_3 and (c) $\rho_2 + \rho_3$ by country, with the •'s at the sample median and the error bars indicating the 90% credible interval. Post-communist countries are in red.

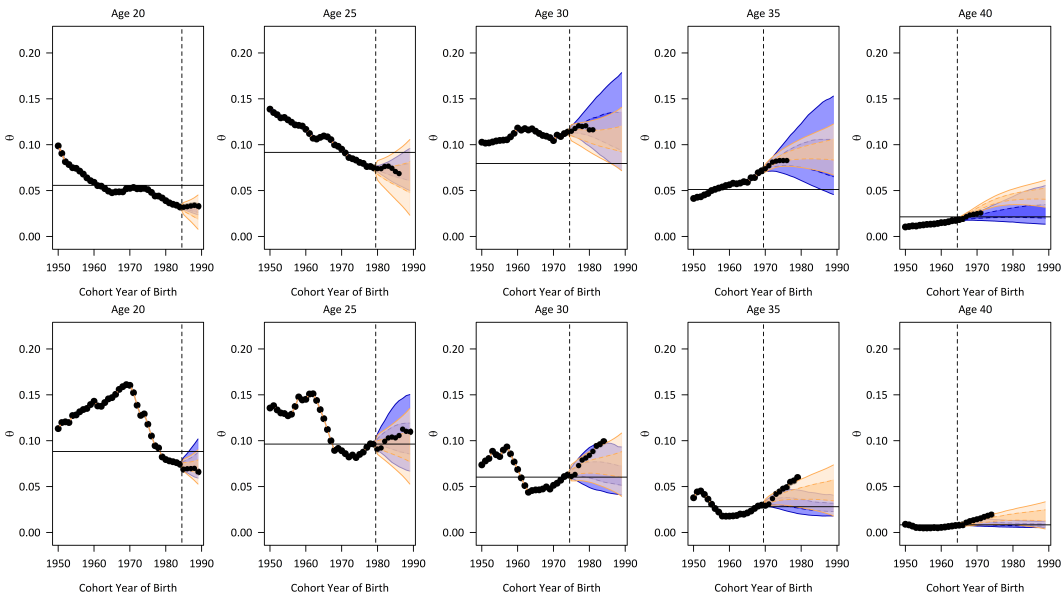

FIGURE 9 2004 Canada (top) and Russia (bottom) fertility forecasts at ages 20, 25, 30, 35 and 40; the dashed line indicates the start of the forecast period and the solid line the median reversion value for the proposed model (Model hB): Model hB 90% credible interval (CI)—; Model hB 50% CI—; the model of Schmertmann et al. (2014a) (Model SZGM) 90% CI—; Model SZGM 50% CI—; Human Fertility Database (2019)—●.

We illustrate this observation by comparing the age-specific forecasts of Canada and Russia, countries with relatively large and small values of this sum respectively, in Figure 9. The Canada forecasts have reasonably wide credible intervals with only slight evidence of reversion to the solid line for the forecast period shown here. Conversely, the Russia forecasts show a very fast reversion and tend to have narrower credible intervals as a result. This provides some

explanation for why our forecasts tend to underperform for the PC countries. It is especially damaging for the Russia forecasts at older ages, where the Model hB intervals are unable to cope sufficiently well with the continued trends that we observe in the forecast period as a result of this reversion. This seems counter-intuitive when we consider that the contemporary data chose to put more weight on following the slope, but instead we have reverted quickly back to the current level. This figure also allows us to assess how successfully we revert back to the current level as imposed by the identifiability constraint (see Section 2.2). Canada and Russia present conflicting results, whereby the former has its median reversion values quite far away from the current level compared to the latter, where they are much closer. This difference could be due to only constraining three linear combinations of the jump-off error terms to equal zero, causing the level of achievement of the desired reversion to vary across countries.

3.3 | 2014 fertility forecasts

Following the 2004 forecasts, we now generate 2014 forecasts for ages 15-44, contemporary cohorts 1960-1999 and historical cohorts 1904-1953. Although some countries have data as recent as 2017 (e.g., Austria and Hungary), others only have data up to 2013 (e.g., Germany and Ukraine). We choose to use the data available in 2014 for these forecasts to ensure that we have 10 or 11 complete contemporary cohorts for each country. We do not use scoring rules to compare the forecasts because there are at most three additional data points for any one country, too few to allow the average to be reliably interpreted. Also, differences in these averages are likely to be negligible due to the minimal uncertainty present when completing the first few cohorts. Furthermore, countries with data up to 2013 or 2014 do not have any observed data to apply a scoring rule to, so we would not be able to compare all countries.

With only qualitative comparison possible, we present the 2014 CFR forecasts for a range of countries in Figure 10. Across these plots we see that, as in Figure 7, the Model hB forecasts tend to be more pessimistic and carry less uncertainty compared to the Model SZGM forecasts. However, we also note that the forecast distributions consistently overlap at least partly across all 30 countries for which we obtained forecasts (see Figure 14 in the online supporting information for the remaining CFR plots not shown in Figure 10). This is somewhat reassuring, as it suggests that the two approaches are able to make roughly similar inferences about the future based on the identical historical and contemporary data that has been fed into them for each country, albeit processed in different ways.

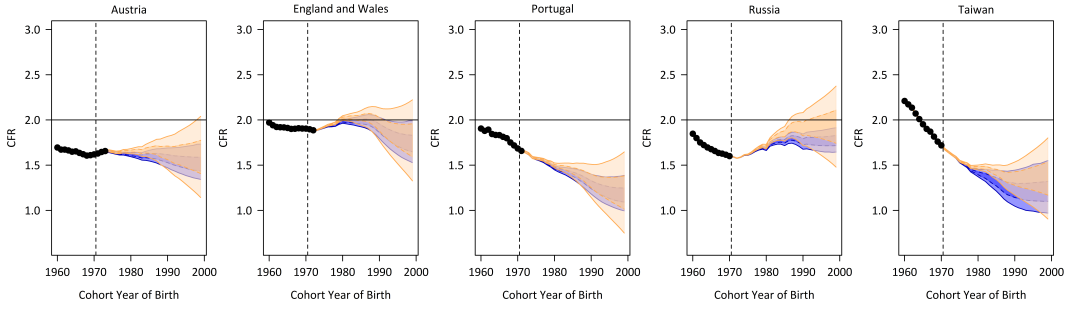

FIGURE 10 2014 cohort total fertility rate (CFR) posterior distributions for selected countries, with the dashed line indicating the start of the forecast period: the proposed model (Model hB) 90% credible interval (CI)—; Model hB 50% CI—; the model of Schmertmann et al. (2014a) (Model SZGM) 90% CI—; Model SZGM 50% CI—; Human Fertility Database (2019)—●.

The first two plots (from L-R) of Austria and England and Wales have the most stationary observed CFR values before the forecast period begins. However whereas the Austria forecast remains without any particular direction, the England and Wales forecast initially shows an increase to replacement level which is then followed by a decline to sub-replacement level for the younger cohorts. The remaining three plots for Portugal, Russia and Taiwan are all forecasting from observed CFR declines in the contemporary Lexis surfaces. Portugal is forecast to continue this decline almost linearly by both models but for Russia the opposite is true, with the forecasts reversing the downward trend; Model SZGM in fact forecasts an optimistic return to replacement level. The subtle difference between the observed CFR values are likely to explain the divergent forecasts for these two countries, namely that Russia shows evidence of the start of an upturn in the CFR just as the forecast period begins whereas Portugal does not. Taiwan is an interesting case whereby like Portugal we see an initial continuation of the decline, but then the downturn stabilises under both models; Model SZGM forecasts this to happen slightly earlier than Model hB.

Regarding stationarity, the analogue of Figure 8 for the 2014 forecasts presented in Figure 11 again shows a preference for the freeze-slope assumption compared to freeze-rate. However whereas for the 2004 forecasts we had no overlap between the error bars in Figures 8a and 8b, here we see substantial overlap for Finland, Iceland and Taiwan. This suggests that ten years after the 2004 forecasts, there is some evidence of a move towards stability in the time series of age-specific fertility rates. Figure 11c demonstrates again not only that all the countries are choosing to follow stationary processes as before, but also that the post-communist countries are still tending to revert faster once the forecast period begins through the comparatively smaller values of $\rho_2 + \rho_3$. However even for a non-post-communist country such as Taiwan, which has a reasonably high level of $\rho_2 + \rho_3$, from the CFR forecast in the right plot of Figure 10 we see clear evidence of the reversion kicking in as we move towards the younger cohorts. This is advantageous as it means that our model will not forecast a trend to continue indefinitely, which would be unrealistic.

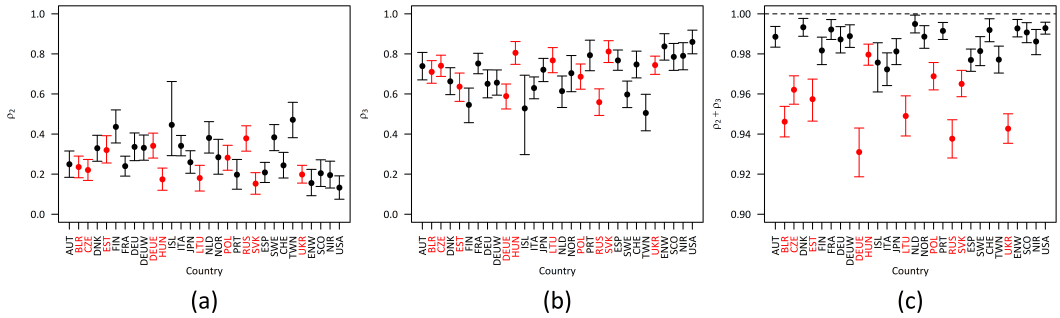

FIGURE 11 2014 posterior distribution summary of (a) ρ_2, (b) ρ_3 and (c) $\rho_2 + \rho_3$ by country, with the •'s at the sample median and the error bars indicating the 90% credible interval. Post-communist countries are in red.

To summarise, even though it is not easy to draw substantive conclusions from the 2014 CFR forecasts due to the lack of validation data available, there is a consistent overlap of the Model hB and SZGM credible intervals across countries. Also, the presence of stationarity in the age-specific forecasts for every country, as in the 2004 forecasts in Section 3.2.3, means that we do not need to be concerned about explosive behaviour in our age-specific or CFR forecasts. Therefore with the little information we have, the 2014 forecasts appear to be plausible and have a well-calibrated level of uncertainty.

4 | DISCUSSION

The aim of this article is to propose a transparent and intuitive hierarchical Bayesian model (Model hB) for forecasting cohort fertility in the spirit of the highly successful model of Schmertmann et al. (2014a), that can compete with the current best cohort fertility forecasting models in terms of forecast accuracy and uncertainty. We incorporate our assumptions, which are similar to those made by Schmertmann et al. (2014a), explicitly into the model structure through a coherent autoregressive time series prior for the error terms (see Section 2.2); the resulting hierarchical form of our model also allows the borrowing of strength across the contemporary cohort-age combinations. The precise specification of the prior is determined by the data, which allows us to learn about the relative weights on staying at the current level (freeze-rate approach) versus following the recent trend (freeze-slope approach) for each country, and subsequently dictating the degree of stationarity of the age-specific processes. The presence of stationarity for all countries in both sets of forecasts makes the reversion level very important - our decision for this level to be as close to jump-off as possible was in the spirit of the recent literature (e.g., see Myrskylä et al. (2013a)). However, this reversion *does* appear to suppress the desire of the data to follow the recent slope in some cases (see Section 3.2.3).

To be considered an important contribution to the literature, it is necessary that our proposed model performs sufficiently well in its purpose, i.e., forecasting age-specific fertility rates and in particular the CFR. To determine this, we carry out an extensive validation of Model hB by comparing its 2004 CFR forecasts for 29 countries against those generated from the models of Schmertmann et al. (2014a), Myrskylä et al. (2013a) and de Beer (1985, 1989), three of the top four models determined by Bohk-Ewald et al. (2018a) in terms of forecast accuracy; in addition to this we compare against the naive freeze rates method, which any justifiable fertility forecasting method should be able to easily outperform. We show that when quantifying the probabilistic accuracy of these forecasts through scoring rules (see Section 3.2.1), there is strong evidence that Model hB is highly competitive in terms of forecast accuracy and uncertainty, as well as unquestionably able to outperform the freeze rates method. The calculation of summary statistics regarding point accuracy and coverage in Section 3.2.2 support these conclusions. For the 2014 forecasts a quantitative comparison is not possible, however the forecasts look reasonable in terms of level and uncertainty (see Section 3.3). So on the whole, Model hB is able to compete well with the current best models in the field. It is important to note, however, that as the Model hB forecasts have only been assessed over a select set of countries at one time point, further validation will be necessary in order to obtain firmer conclusions regarding forecast performance.

A key advantage of the competing models is their low computational cost, as it takes seconds to produce a fit for one country. Model hB requires MCMC methodology and therefore large numbers of iterations are sometimes necessary to obtain results of a suitable quality. This can be computationally expensive; however, thanks to the state-of-the-art Hamiltonian Monte Carlo methods and the RStan software package (Stan Development Team, 2018a), posterior sampling can be conducted with reasonable efficiency. To quantify this, we found the average fitting time per country for the 2004 and 2014 forecasts to be 15 and 21 minutes respectively, with no country taking longer than one hour to fit; this is not an unreasonable length of time in practice, especially if it makes the underlying model assumptions more realistic and provides adequate levels of uncertainty. Regarding the simple extrapolation method of Myrskylä et al. (2013a), it may perform well in terms of forecast accuracy but does not have well-calibrated levels of uncertainty (see Section 3.2.2); the model of de Beer (1985, 1989) and the freeze rates method do not provide any uncertainty quantification and therefore can only have limited use as deterministic forecasting approaches. Hence, we believe that the use of complex statistical methods in cohort fertility forecasting models such as Model hB and the conjugate Bayesian model of Schmertmann et al. (2014a) *is* worth the effort, in response to the question posed

by Bohk-Ewald et al. (2018a). This is especially important if the aim is to obtain long-term fertility forecasts, as long time series of rates are necessary in order to have appropriate uncertainty.

The finding regarding the greater success in forecasting for non-post-communist countries deserves some discussion. Clearly the contemporary Lexis surfaces of post-communist countries provide a stronger forecasting challenge due to the sharp decline in the time series of age-specific fertility rates following the regime change. Model hB seems to respond to this by decreasing the combined sum of the weights on the freeze-rate and freeze-slope approaches, leading to an increased degree of stationarity and therefore a faster reversion in the forecast period (see Section 3.2.3). This suggests that the sensitivity of Model hB to recent data could be a disadvantage when there has been a recent structural shift in the country of interest. In an attempt to decrease this sensitivity, we experimented with allowing ρ_2 and ρ_3 to borrow strength across countries; however, the post-communist countries appeared only able to tolerate very slight increases to the value of $\rho_2 + \rho_3$, which were insufficient to noticeably influence the forecasts. Further investigation into the reasons behind this inconsistency in performance, as well as its reduction, will be required.

As mentioned earlier, further validation of Model hB needs to be carried out. This should involve expanding the set of countries considered so that we can assess performance in a wider range of circumstances, for example high fertility settings. Additionally, multiple forecasts should be generated at various time points to allow quantitative assessment of the performance of Model hB across time and in the long term. We also aim to make further attempts to improve the forecast performance of Model hB for countries exhibiting a recent structural break. To this end, potential avenues to explore are restricting the contemporary data used for such countries, incorporating expert opinion, and imposing a constraint on the shape of the cohort schedules; this latter suggestion is the one key assumption of the model of Schmertmann et al. (2014a) that we did not build into our proposed model. The superior performance of the model of Myrskylä et al. (2013a) in terms of forecast point accuracy (see Section 3.2.2) provides evidence that such constraints are not vital in order to achieve reasonable success; however, it could still be useful to investigate, especially the implications on long-term forecasts. Given the results from this paper, when performing fertility forecasting we recommend fitting a selection of models that have been shown to be competitive in the literature, and fully exploring the causes of any divergences occurring among the forecasts.

In conclusion, our hierarchical Bayesian approach to forecasting cohort fertility is successful through its transparent specification and competitive forecast performance when compared against three of the current best models in the field according to Bohk-Ewald et al. (2018a), in particular for countries without a recent structural shift. In addition, it demonstrates how advanced computational methods can be used to fit hierarchical Bayesian models with an atypical setup. This not only cements the position of hierarchical Bayesian methods at the forefront of population forecasting methods, but also makes a valuable contribution to the fertility modelling and forecasting literature.

ACKNOWLEDGEMENTS

The PhD programme of the first author is funded by the EPSRC (award reference 1801045). The work of the second and third authors is partly supported by the ESRC Centre for Population Change - phase II (grant ES/K007394/1). The authors would like to thank Jakub Bijak, Jason Hilton, and Peter Smith for their feedback during the initial research and writing phases of the project. The authors also acknowledge Carl Schmertmann and anonymous reviewers, who provided helpful comments on earlier versions of this paper.

REFERENCES

de Beer, J. (1985) A time series model for cohort data. *Journal of the American Statistical Association*, **80**, 525–530.

— (1989) Projecting age-specific fertility rates by using time-series methods. *European Journal of Population*, **5**, 315–346.

Bijak, J. and Bryant, J. (2016) Bayesian demography 250 years after Bayes. *Popln Stud.*, **70**, 1–19.

Billingsley, S. (2010) The Post-communist Fertility Puzzle. *Population Research and Policy Review*, **29**, 193–231.

Bohk-Ewald, C., Li, P. and Myrskylä, M. (2018a) Forecast accuracy hardly improves with method complexity when completing cohort fertility. *Proceedings of the National Academy of Sciences*, **115**, 9187–9192.

— (2018b) *Forecast accuracy hardly improves with method complexity when completing cohort fertility* [R code]. URL: `https://github.com/fertility-forecasting/validate-forecast-methods`. Accessed 4 January 2019.

Bongaarts, J. and Feeney, G. (1998) On the quantum and tempo of fertility. *Population and development review*, 271–291.

Booth, H. (2006) Demographic forecasting: 1980 to 2005 in review. *Int. J. Forecast.*, **22**, 547–581.

Brass, W. (1960) The graduation of fertility distributions by polynomial functions. *Popln Stud.*, **14**, 148–162.

— (1974) Perspectives in Population Prediction: Illustrated by the Statistics of England and Wales. *J. R. Statist. Soc.* A, **137**, 532–583.

Bröcker, J. (2012) Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, **138**, 1611–1617.

Chandola, T., Coleman, D. A. and Hiorns, R. W. (1999) Recent European fertility patterns: Fitting curves to 'distorted' distributions. *Popln Stud.*, **53**, 317–329.

Coale, A. J. and Trussell, T. J. (1974) Model fertility schedules: Variations in the age structure of childbearing in populations. *Population Index*, **40**, 185–258.

Czado, C., Delwarde, A. and Denuit, M. (2005) Bayesian Poisson log-bilinear mortality projections. *Insur. Math. Econ.*, **36**, 260–284.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014) *Bayesian Data Analysis*. Florida: Chapman & Hall/CRC, third edn.

Girosi, F. and King, G. (2008) *Demographic Forecasting*. Princeton, New Jersey: Princeton University Press.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Good, I. J. (1952) Rational decisions. *J. R. Statist. Soc.* B, **14**, 107–114.

Hadwiger, H. (1940) Eine analytische Reproduktionsfunktion für biologische Gesamtheiten [An analytic reproduction function for biological groups]. *Skandinavisk Aktuarietidskrift [Scand. Act. J.]*, **23**, 101–113.

Hoem, J. M., Madsen, D., Nielsen, J. L., Ohlsen, E., Hansen, H. O. and Rennermalm, B. (1981) Experiments in modelling recent Danish fertility curves. *Demography*, **18**, 231–244.

Human Fertility Database (2019) Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). URL: `http://www.humanfertility.org`. Data downloaded 15 February 2019.

Hyndman, R. J. and Ullah, M. S. (2007) Robust forecasting of mortality and fertility rates: A functional data approach. *Computnl Statist. Data Anal.*, **51**, 4942–4956.

Jasilioniene, A., Jdanov, D. A., Sobotka, T., Andreev, E. M., Zeman, K., Shkolnikov, V. M., with contributions from Goldstein, J., Nash, E. J., Philipov, D. and Rodriguez, G. (2015) *Methods Protocol for the Human Fertility Database*. URL: `http://www.humanfertility.org/Docs/methods.pdf`. Accessed 27 April 2017.

Jordan, A., Krüger, F. and Lerch, S. (2017) Evaluating probabilistic forecasts with scoringRules. *arXiv e-prints*, arXiv:1709.04743. Accessed 16 January 2018.

Krüger, F., Lerch, S., Thorarinsdottir, T. L. and Gneiting, T. (2019) Predictive Inference Based on Markov Chain Monte Carlo Output. *arXiv e-prints*, arXiv:1608.06802. Accessed 27 November 2019.

Lee, R. D. (1992) Stochastic demographic forecasting. *Int. J. Forecast.*, **8**, 315–327.

Li, N. and Wu, Z. (2003) Forecasting cohort incomplete fertility: A method and an application. *Population Studies*, **57**, 303–320.

Matheson, J. E. and Winkler, R. L. (1976) Scoring rules for continuous probability distributions. *Management Science*, **22**, 1087–1096.

Myrskylä, M., Goldstein, J. R. and Cheng, Y. A. (2013a) New cohort fertility forecasts for the developed world: Rises, falls, and reversals. *Population and Development Review*, **39**, 31–56.

— (2013b) *New Cohort Fertility Forecasts for the Developed World: Rises, Falls, and Reversals* [Stata code]. URL: `https://www.demogr.mpg.de/go/cohort_fertility/`. Accessed 4 January 2019.

National Records of Scotland (2019) Uses and limitations of population projections. `https://tinyurl.com/ub6hlfz`. Accessed 13 November 2019.

Ní Bhrolcháin, M. (2011) Tempo and the TFR. *Demography*, **48**, 841–861.

Office for National Statistics (2017) *National Population Projections Consultation - 2016-based national population projections: fertility*. `https://tinyurl.com/y23dzv4a`. Accessed 20 February 2019.

— (2019a) National Population Projections Quality and Methodology Information (QMI). `https://tinyurl.com/ybj58awe`. Accessed 13 November 2019.

— (2019b) *Births Quality and Methodology Information (QMI)*. `https://tinyurl.com/yawcgzya`. Accessed 13 November 2019.

Peristera, P. and Kostaki, A. (2007) Modeling fertility in modern populations. *Demog. Res.*, **16**, 141–194.

Population Reference Bureau (2001) Understanding and Using Population Projections. `https://www.prb.org/wp-content/uploads/2001/12/UnderStndPopProj_Eng.pdf`. Accessed 13 November 2019.

Schmertmann, C., Zagheni, E., Goldstein, J. R. and Myrskylä, M. (2014a) Bayesian Forecasting of Cohort Fertility. *J. Am. Statist. Ass.*, **109**, 500–513.

— (2014b) *Bayesian Forecasting of Cohort Fertility* [project website]. URL: `http://schmert.net/cohort-fertility/`. Accessed 16 February 2018.

Ševčíková, H., Li, N., Kantorová, V., Gerland, P. and Raftery, A. E. (2016) Age-specific mortality and fertility rates for probabilistic population projections. In *Dynamic Demographic Analysis* (ed. R. Schoen), 285–310. Cham, Switzerland: Springer International Publishing.

Shang, H. L. (2012) Point and interval forecasts of age-specific fertility rates: a comparison of functional principal component methods. *J. Popln Res.*, **29**, 249–267.

Stan Development Team (2018a) *RStan: The R interface to Stan*. R package version 2.17.4. URL: `http://mc-stan.org/`.

— (2018b) *Stan Modeling Language: User's Guide and Reference Manual*. Version 2.18.0. URL: `http://mc-stan.org/`.

Wiśniowski, A., Smith, P. W. F., Bijak, J., Raymer, J. and Forster, J. J. (2015) Bayesian Population Forecasting: Extending the Lee-Carter Method. *Demography*, **52**, 1035–1059.

A | QUANTIFYING UNCERTAINTY FOR A CONJUGATE MODEL

The posterior distribution for the conjugate model of Schmertmann et al. (2014a) is $(\theta|y_{\text{country}}) \sim N(\mu_{\text{post}}, \Sigma_{\text{post}})$, using the same notation as the authors. To incorporate the additional variation described in Section 2.3, we add a modified version of $\boldsymbol{\Psi}$, the covariance matrix for the observed rates, to Σ_{post}. We denote this by $\boldsymbol{\Psi}^* := \text{diag}_{j=1,\ldots,CA}\left[[\mu_{\text{post}}]_j / W_j^*\right]$. We modify $\boldsymbol{\Psi}$ by first extending it to all CA cohort-age combinations, with index j corresponding to the jth combination when ordered by age within cohort. We then evaluate the numerator of each entry at its corresponding value of μ_{post} rather than y, but the denominator W_j^* remains as the jth exposure; in the same spirit as Section 2.3, W_j^* is taken to be its most recently observed value at age a if unobserved. We then compute the 90% and 50% credible intervals for the empirical birth rates as $\mu_{\text{post}} \pm z\sqrt{\text{diag}(\Sigma_{\text{post}} + \boldsymbol{\Psi}^*)}$, where $z \approx 1.64$ and 0.67 respectively. This will only have a noticeable effect for small countries with comparatively low exposures for some cohort-age combinations.

B | DISTRIBUTION OF THE AVERAGE CONTINUOUS RANKED PROBABILITY SCORE ACROSS COUNTRIES FOR EACH MODEL

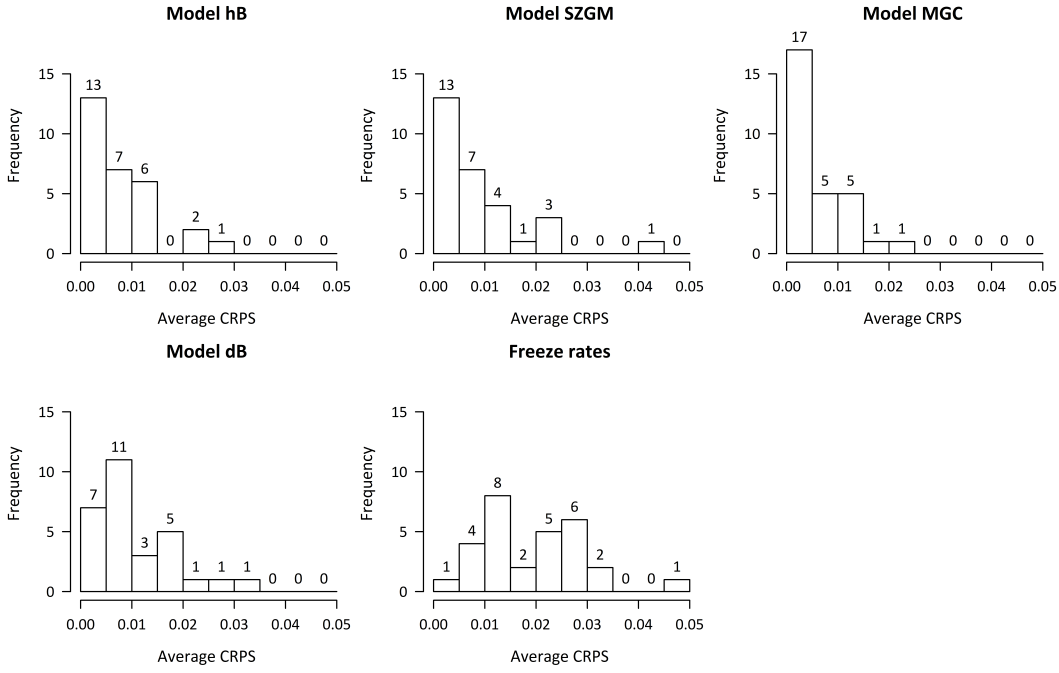

FIGURE 12 Histograms of the average continuous ranked probability score (CRPS) across countries for the 2004 cohort total fertility rate (CFR) forecasts for the proposed model (hB), the models of Schmertmann et al. (2014a) (SZGM), Myrskylä et al. (2013a) (MGC) and de Beer (1985, 1989) (dB), and the naive freeze rates approach.