Can Incorporating Parity Information Improve the Reliability of Completed Cohort Fertility Projections? Insights from a Bayesian Generalized Additive Model Approach

Joanne Ellison[1], Jakub Bijak[2], Erengul Dodd[3]

[1,2]Department of Social Statistics and Demography, University of Southampton, UK

[3]School of Mathematical Sciences, University of Southampton, UK

[1]J.V.Ellison@soton.ac.uk; [2]J.Bijak@soton.ac.uk; [3]E.Dodd@soton.ac.uk

Abstract

Fertility projections inform population projections and are used to plan for the future provision of vital services such as maternity care and schooling. Existing fertility forecasting models tend to use aggregate births data indexed by age and time alone, thereby neglecting to include information about parity, i.e. the number of previous live-born children. This omission risks ignoring a crucial mechanism of fertility dynamics. We propose a Bayesian parity-specific fertility projection model to complete cohort fertility, within a generalized additive model (GAM) framework. The use of GAMs enables a smooth age-cohort rate surface to be estimated for each parity simultaneously. We constrain our model using aggregate data and additionally introduce random walk priors on completed family size and parity progression ratios, which are summary fertility measures known to change relatively slowly over time. Using Hamiltonian Monte Carlo methods and data from the Human Fertility Database, we fit our model to 16 countries. We compare our forecasts with the best-performing existing models to quantify the impact of including the parity dimension on predictive accuracy. Our findings indicate that a parity-specific approach could lead to more plausible and reliable fertility projections, aiding government planners in their decision-making and enabling more tailored policy solutions.

Keywords: Bayesian methods; Cohort fertility; Forecasting; Generalized additive models; Parity progression

1 Introduction

Fertility projections are used in population projections and to plan for future service provision, such as maternity care and schooling. Methodological literature on fertility forecasting is extensive, with comprehensive reviews in Booth (2006) and Bohk-Ewald et al. (2018a). Existing models tend to use aggregate data on the total number of births indexed only by age, and period (calendar year) or birth cohort. These variables constitute three of the "five clocks" of fertility (Raftery et al. 1996:1), the other two being time since last birth and number of previous live-born children (parity). While little information is typically collected about time since last birth at the population level, disaggregation of births by birth order is increasingly common (Jasilioniene et al. 2015). As childbearing is a sequential process, the current number of children is an important determinant of the decision to have another child; therefore, omitting birth order information risks ignoring a crucial mechanism of fertility. A parity-specific approach can thus translate into more reliable projections and improved decision-making in a range of policy and planning domains, while also producing forecasts of parity-specific measures which are themselves of great interest.

While the consideration of parity is underexplored in fertility projections (Petropoulos et al. 2022:797–798), it is commonplace in the fertility analysis literature. A key reason for this is the ability to gain a more detailed understanding of the changes in childbearing behaviour underlying past trends in fertility levels. The approaches taken include: computing period parity progression ratios (Feeney and Yu 1987; McDonald et al. 2015; Murphy and Berrington 1993; Ní Bhrolcháin 1987); proposing parity-adjusted summary fertility measures such as the total fertility rate (Bongaarts and Sobotka 2012); decomposing changes in parity progression ratios and total fertility rates by parity (Andreev et al. 2002; Hellstrand et al. 2020, 2021) and changes in the cohort total fertility rate by parity progression ratios (Zeman et al. 2018); and calculating standardized rates by birth order (Andersson 1999; Rallu and Toulemon 1994). Event history analyses of births also typically account for parity, by fitting separate parity-specific models (e.g. Ellison et al. 2022) or joint models that control for parity and unobserved heterogeneity between women (e.g. Begall and Mills 2013; Kravdal 2001;

Van Hook and Altman 2013).

Analysing childbearing patterns by parity can also better inform future fertility assumptions (Smallwood 2002), which are needed when producing projections. Indeed, Kohler and Ortega (2002) argue that appropriate cohort fertility projections require plausible assumptions for the quantum (level) and tempo (timing) of fertility, which in turn necessitate parity information to account for those exposed to each birth order more precisely; concurrently, this captures the aforementioned sequential nature of the childbearing decision-making process. This additional differentiation by parity increases the homogeneity of the resulting groups while also incorporating women's previous childbearing experience (Rallu and Toulemon 1994). Hobcraft (1996) provides a historical example, stating that parity-specific analyses would likely have helped to anticipate the baby boom and bust of the 1960s and 1970s in England & Wales, which were not predicted by the corresponding forecasts. Regarding current use, a recent European survey of 32 national statistical agencies found that while eight use parity-specific data to inform their fertility projections, only two actually produce parity-specific projections (Gleditsch et al. 2021).

The core of our proposal is that incorporating a lower level of disaggregation into the forecasting model could improve predictive accuracy at the aggregate level. This is closely related to the "bottom-up" approach, well-known in the broader forecasting literature, whereby forecasts are obtained at a lower level and subsequently aggregated (Petropoulos et al. 2022). This approach has been applied in a range of contexts in the social sciences, including forecasting economic indicators (e.g. Duarte and Rua 2007; Heinisch and Scheufele 2018), and mortality rates by cause of death (Li et al. 2019) and subnational area (Li and Hyndman 2021). Bottom-up methods are typically compared with "top-down" approaches, which forecast at the highest level of aggregation and then disaggregate (Petropoulos et al. 2022). Considerable support has been found for the bottom-up method over the top-down approach (Athanasopoulos et al. 2024), however we are more concerned with the choice of disaggregation level. Various studies have explored this issue, including in the economic (Duarte and Rua 2007; Zotteri et al. 2005) and demographic (Ahlburg 2001; Lutz et al. 1998) forecasting literature. In particular, Lutz et al. (1998) specify that a

proposed additional dimension should be of substantive interest, a source of demographic heterogeneity, and technically possible to incorporate. Furthermore, Ahlburg (2001) comments that although disaggregation may not improve predictive performance at the aggregate level, there are other benefits such as the generation of knowledge and accurate subpopulation forecasts.

In the fertility projections literature, the production of parity-specific forecasts has been suggested, for example by Schmertmann et al. (2014a) and Shang and Booth (2020), with data limitations preventing the latter from implementing the approach. Ellison et al. (2023) propose a Bayesian parity-specific projection model combining survey and vital registration data. Aggregate forecasts are generated but each parity is modelled independently – this means that measures depending on multiple parities, such as age-specific fertility rates (ASFRs) or summary measures, cannot be constrained. The authors recommend the subsequent development of a generalizable approach that models the parities jointly and incorporates interpretable constraints on summary measures. A Bayesian setup enables coherent integration of multiple data sources and prior knowledge, and appropriate quantification of predictive uncertainty; in the field of demographic forecasting, Bayesian methods are becoming increasingly prevalent (Mazzuco and Keilman 2020).

This discussion motivates the aims of this paper. Firstly, we aim to develop a fertility forecasting method to complete cohort fertility that incorporates parity information, advancing the existing literature in line with the suggestions of Ellison et al. (2023). Secondly, we aim to assess whether the proposed method has increased forecast reliability. Within a Bayesian generalized additive model framework and informed by parity-specific rate estimates from the Human Fertility Database (HFD) (HFD 2024), we construct smooth parity-specific age-cohort rate surfaces. We additionally incorporate constraints on ASFRs and summary fertility measures at the parity-specific and aggregate levels, using pooled data also from HFD (2024). We obtain forecasts for 16 countries, comparing their performance with some of the best-performing existing models and an aggregate version of our proposed model.

The paper proceeds as follows. After describing our data sources in Section 2, we specify our modelling approach in Section 3. In Section 4 we present the model outputs, namely the parity-specific rates and summary measures, and summarize the comparative findings. We then provide a discussion in Section 5.

2 Data

2.1 Setup

In this paper we address the fertility projection problem of completing cohort fertility: forecasting the average number of children that women from a particular birth cohort will have had by the end of their reproductive lives (cohort total fertility rate, CFR). Contrastingly, period fertility considers the average number of children a pseudo-cohort of women would have if they were subject to the fertility rates of a given period at every age (total fertility rate, TFR). The pros and cons of these approaches to fertility analysis are well known (Ní Bhrolcháin 1992; Schoen 2022), and either could be used in this context; however, we prefer the cohort approach for several reasons. Firstly, where period data on parity distributions are absent or lacking in quality and/or frequency, a cohort perspective allows their estimation by cumulating across each reproductive age (Jasilioniene et al. 2015; also see Section 2.2). Secondly, the reduced fluctuation of cohort fertility (de Beer 1985; Li and Wu 2003; Sobotka 2003) is beneficial for forecasting, as it justifies more strongly the assumption of the future smooth progression of summary measures.

Next we introduce the modelling setup. Taking the reproductive age range to be from $a_{min} = 15$ to $a_{max} = 44$ years, we let JOY be a particular 'jump-off' year whereby its associated data are the most recent data used in model fitting. We consider four JOY values – 2000, 2005, 2010 and 2015 – to test the forecast performance of our model over different periods and forecast horizons. In a similar setup to Schmertmann et al. (2014a), for a given JOY we consider $n_c = a_{max} - a_{min} + 11 = 40$ cohort years of birth, ranging from $c_{min}^{JOY} = (JOY - a_{max} - 10)$ to $c_{max}^{JOY} = (JOY - a_{min})$. For example, $c_{min}^{2000} = 1946$ and $c_{max}^{2000} = 1985$, while $c_{min}^{2015} = 1961$ and $c_{max}^{2015} = 2000$. The first 11 cohorts (the 1961–1971 cohorts

for $JOY = 2015$) have reached at least age a_{max} by year JOY and therefore have already completed their reproductive lives; the subsequent $(a_{max} - a_{min})$ cohorts are only partially observed up to age $(a_{max} - 1), \ldots, a_{min}$. We illustrate this setup graphically in panel a of Figure 1, for the Canada ASFRs when $JOY = 2015$. The rates are displayed in the form of a Lexis surface (left plot) and cohort-specific rate curves (right plot).

The data requirements of the model are population-level birth counts (by age, cohort and birth order), and exposures (by age and cohort). We use HFD estimates of age-cohort births (by order) and exposures, computing exposure estimates corresponding to age-cohort combinations that are either observed (but not available in the HFD) or in the future, from the 2024 Revision of *World Population Prospects* (WPP) (United Nations, Department of Economic and Social Affairs, Population Division 2024). The WPP population estimates and projections are given by country, sex and single year of age on January 1st each year – we first approximate each female age-period exposure by the corresponding mid-year population estimate or projection, and then convert these values to age-cohort exposure estimates via Lexis triangles, as recommended by van Raalte et al. (2023) when working with period data.

Across the four JOY values we fit our model to 16 HFD countries, listed in Appendix A. Although many more countries have order-specific births data available in the HFD, it is not possible to include them either because the time series is not long enough (the parity-specific data often start to be collected much later than the aggregate data) and/or the data are not of sufficiently high quality. To include as many countries as possible, we allow any subset of the JOY values to be represented for a given country; we also allow for a "reduced" setup of the model, with $n_c = 35$ (instead of 40) cohorts considered and only the first 6 (instead of 11) being fully observed (see Section 2.1). The 16 countries included exhibit considerable geographical spread, and differing population sizes and social, cultural and historical contexts. This leads to a diverse set of childbearing patterns upon which we can suitably test the forecast performance of our model. The supplementary material contains a spreadsheet which describes the available data for each HFD country, summarizes any relevant data quality issues mentioned in the HFD documentation, and states the subsequent decision taken (to include all or a subset of the cohorts for the country,

or to exclude it completely).

2.2 Age-parity- and age-specific fertility rates

The primary data source that informs our fertility projection model is conditional parity-specific fertility rate estimates by age and cohort. Their conditionality means that they are of occurrence/exposure form and so appropriately account for the population at risk. For women of a given age a and cohort c, we define the conditional fertility rate for birth order i, l^i_{ac}, in the following way:

$$l^i_{ac} = \frac{\text{births of order } i \text{ to women aged } a \text{ in cohort } c}{\text{person-years lived at parity } i-1 \text{ by women aged } a \text{ in cohort } c} = \frac{b^i_{ac}}{E^{i-1}_{ac}}. \tag{1}$$

We assume that parities J and above, and birth orders $J+1$ and above, are combined. Therefore parity takes values $0, 1, \ldots, J+$ and birth order takes values $1, 2, \ldots, (J+1)+$. We take $J = 3$ and therefore consider parities 0–3+, due to small cell counts and similar age patterns for the conditional rates of the highest parities.

Data for the numerator comes from HFD (2024) (see Section 2.1). Information on true (i.e. biological) birth order is now collected by most countries, with many having only requested parity information for marital births until relatively recently (Jasilioniene et al. 2015). We estimate the denominators by cumulating cohort fertility, following the method of Smallwood (2002); the HFD produces its fertility tables using a similar approach (Jasilioniene et al. 2015). The method first assumes that all women in a given cohort have no children at age a_{min}. Then, at each subsequent age, the relevant order-specific birth counts are used to update the number of women exposed to each parity. These are then scaled to sum to the corresponding exposure estimate, giving the final parity-specific exposures. We specify the method formally in Appendix B.

A drawback to this approach is that it ignores any potential sensitivity of parity to mortality and migration, which would ideally be accounted for if data were available (Smallwood 2002). We also note that although the HFD provides census- or register-based estimates of the female age-parity distribution for the vast majority of the 16 countries we consider, we

do not make use of this in the main analysis[1] for several reasons. Firstly, annual estimates are only available for the Nordic countries from their population registers[2], with estimates for most of the other countries only available from the decennial census; this infrequency makes it difficult to incorporate the data into the cohort cumulation process. The HFD achieves this by using a single census or register estimate for the starting exposures of its period fertility tables, when it is possible and appropriate (Jasilioniene et al. 2015); however, this limits the range of years for inclusion. Secondly, the absence of such estimates for some countries, and the differing data availability for the others, means that any proposed approach would be inconsistently applied across countries. Thirdly, it will be necessary to replicate the chosen estimation method within our projection model, further elevating the importance of a simple and consistent methodology.

Next, for women of a given age a and cohort c, we define the unconditional fertility rate for birth order i, m^i_{ac}, as:

$$m^i_{ac} = \frac{\text{births of order } i \text{ to women aged } a \text{ in cohort } c}{\text{person-years lived by women aged } a \text{ in cohort } c} = \frac{b^i_{ac}}{E_{ac}}. \tag{2}$$

The only difference from the conditional rate in equation (1) is the change in denominator from women at risk of the particular birth order to those at risk of any birth.

Lastly, for women of a given age a and cohort c, we define the ASFR or age-specific fertility rate, m_{ac}, as:

$$m_{ac} = \frac{\text{births to women aged } a \text{ in cohort } c}{\text{person-years lived by women aged } a \text{ in cohort } c} = \frac{\sum_i b^i_{ac}}{E_{ac}} = \frac{b_{ac}}{E_{ac}}.$$

We note that $m_{ac} = \sum_i m^i_{ac}$, i.e. for fixed a and c, the sum of the unconditional rates gives the ASFR.

[1] In Appendix C we compare the census- or register-based estimates with our cumulation-based estimates, finding that the two methods correspond reasonably closely with each other. Also, in Appendix L we investigate the sensitivity of our forecasts to the two types of exposures; we discuss this further in Section 5.

[2] Hungary also has annual estimates available, but these interpolate between decennial censuses rather than being independent annual estimates.

2.3 Summary fertility measures

We describe the computation of the summary fertility measures required for our proposed model, namely the CFR (see Section 2.1) and parity progression ratios (PPRs). These are defined for women in cohort c after age a_{max}. We define the CFR, CFR_c, representing the average completed family size, as $CFR_c = \sum_{a_{min}}^{a_{max}} m_{ac}$. We define the CFR for birth order i, CFR_c^i, representing the average number of births of order i per woman, as $CFR_c^i = \sum_{a_{min}}^{a_{max}} m_{ac}^i$. We note that $CFR_c = \sum_i CFR_c^i$. Then we define the PPR from parity i to $i+1$, $PPR_c^{i\to i+1}$, representing the proportion of women progressing from parity i to $i+1$, as:

$$
\begin{aligned}
PPR_c^{0\to 1} &= CFR_c^1; \\
PPR_c^{i\to i+1} &= \frac{CFR_c^{i+1}}{CFR_c^i} \text{ for } i \in \{1,2,\ldots,(J-1)\}; \\
PPR_c^{J+\to (J+1)+} &= \frac{CFR_c^{(J+1)+}}{CFR_c^J + CFR_c^{(J+1)+}}.
\end{aligned}
$$

For a given jump-off year JOY, using estimates of the m_{ac}^i values from the HFD, we calculate these summary measures for all country-cohort combinations from the 1945 cohort onwards who have completed their reproductive lives by year JOY and where the estimates are deemed to be of sufficiently high quality. We use these pooled datasets to inform constraints that we incorporate into our proposed model (Section 3.3 provides further details). Having described the model setup, the required data sources and fertility measures, in Section 3 we specify our parity-specific projection model.

3 Methods

3.1 Model structure

We illustrate the model structure in Figure 2. We distinguish between three levels – Level-1, Level-2 and Level-3 – in the columns. Each level has a different demographic array of interest (white rectangles):

- Level-1: true conditional birth rates by age, cohort and birth order, i.e. the true l^i_{ac} values (equation (1) in Section 2.2);
- Level-2: true unconditional birth rates by age, cohort and birth order, i.e. the true m^i_{ac} values (equation (2) in Section 2.2);
- Level-3: true summary fertility measures, namely the true CFR and PPRs (Section 2.3).

The observed quantities (gray rectangles) refer to the data requirements of the model (Sections 2.1–2.3):

- birth counts by age, cohort and birth order (from the HFD; Level-1), which then sum to give birth counts by age and cohort (Level-2);
- exposures by age and cohort (from the HFD and WPP; Level-2), and exposures by age, cohort and parity (approximated using the cumulation-based approach described in Section 2.2; Level-1);
- pooled cross-country dataset of CFR and PPR estimates from the HFD (Level-3).

The two system models (rounded rectangles) allow us to incorporate assumptions regarding the progression of the demographic arrays of interest across age and/or time. The first (Level-1) assumes underlying smoothness of the true conditional order-specific birth rates across age and cohort for each birth order. The second (Level-3) assumes slow changes of the CFR and PPRs over time informed by the HFD data. The two data models (hexagons) describe the processes through which the birth count data are generated from the true rates. These govern the relationships between the births by age, cohort and birth order and the true conditional birth rates (Level-1), and the births by age and cohort and the true unconditional rates (Level-2). We compute the true unconditional rates from the true conditional rates by again applying the cumulation-based method to generate order-specific birth counts according to the true conditional rates, which we then divide by the age-cohort exposures which are assumed known. The summary measures then follow straightforwardly from the formulas in Section 2.3. The remainder of this section motivates and specifies the system and data models corresponding to Level-1 and Level-2 (Section 3.2)

and Level-3 (Section 3.3).

3.2 Modelling rates by age, cohort and parity (Level-1 and Level-2)

3.2.1 System model

First we specify the Level-1 system model for the true conditional birth rates by age, cohort and birth order. We will interchangeably refer to these as parity-specific rates for current parities 0, 1, 2 and 3+, and order-specific rates for progression to first, second, third, or fourth and higher-order birth. Returning to the age-cohort ASFR surface (panel a of Figure 1), we now consider four such surfaces of the conditional parity-specific rates, illustrated in panel b. Analogous to panel a, we present the Lexis surfaces and the cohort-specific rate curves on the top and bottom rows respectively. While the first birth (parity 0) rates progress smoothly across age and cohort and display similar patterns to the ASFRs, the higher-order rates are more erratic, particularly at the youngest ages[3].

Our approach smooths these rates across age and cohort for each parity on the logarithmic scale, within a generalized additive model (GAM) framework. GAMs enable individual and joint covariate effects to be represented as smooth functions; Wood (2017) provides an in-depth introduction. In the demographic forecasting literature, GAMs have been applied to mortality. Hilton et al. (2019, 2021) develop Bayesian approaches estimating smooth functions of age, age-specific improvement factors, and cohort. Aburto et al. (2021) fit a GAM including smooth effects of age, week and their interaction to forecast weekly deaths, finding it to outperform existing methods. Currie et al. (2004) perform two-dimensional (2D) smoothing of mortality rates across age and period under a P-spline approach (Eilers and Marx 1996). Camarda (2019) extends such an approach to incorporate shape constraints and thus improve forecast plausibility. Concerning fertility, the Bayesian forecasting model of Ellison et al. (2023) (see Section 1) is also based on a GAM framework. The model estimates smooth effects of age, cohort and their interaction, as well as time since last birth.

[3]This observation is the result of several factors: few women are exposed, and those that have reached parity 1 or 2 as a teenager are a select group who have a reasonably high likelihood of progressing quickly to higher parities (Greulich and Toulemon 2023). It is also the case that our cumulation-based denominators are likely to be an underestimate of the true exposure (see Appendix C), leading to even greater fluctuation in the rates.

We now specify our system model. Letting λ^i_{ac} be the true conditional rate for age a, cohort c and birth order i, our system model sets:

$$\log\left(\lambda^i_{ac}\right) = f_i(a,c),$$

where f_i is a smooth function of age and cohort estimated using the observed data. Similarly to Currie et al. (2004) and Camarda (2019), we take a P-spline approach. In the simple one-dimensional (1D) case this considers a cubic B-spline basis, consisting of a series of smooth curves, or basis functions, that cover the covariate range and are non-zero for different parts of it. Following the advice of Currie et al. (2004) we space the basis functions regularly at five-year intervals, requiring eight and ten in the age and cohort dimensions respectively. The smooth curve is then constructed as a linear combination of the basis functions, where the coefficients are constrained to change smoothly across the covariate. Appendix D provides a visualization of this 1D case.

Extending this to smooth the log-rates across age *and* cohort, the basis functions are now smooth 2D surfaces constructed by multiplying the curves from the age and cohort dimensions together[4]. Letting there be $P = 8$ and $Q = 10$ 1D basis functions covering the age and cohort ranges respectively, for a given age a and cohort c we represent the values of these functions by $b^1_A(a),\ldots,b^P_A(a)$ and $b^1_C(c),\ldots,b^Q_C(c)$. Then, for a given $p \in \{1,\ldots,P\}$ and $q \in \{1,\ldots,Q\}$ we represent the value of the corresponding 2D basis function as $b^{pq}_{AC}(a,c)$, where $b^{pq}_{AC}(a,c) = b^p_A(a) \times b^q_C(c)$. We then express $f_i(a,c)$ as the weighted sum of these basis functions with parity-specific weights β^{pq}_i:

$$f_i(a,c) = \sum_{p=1}^{P}\sum_{q=1}^{Q} \beta^{pq}_i b^{pq}_{AC}(a,c).$$

In our case we therefore have $P \times Q = 8 \times 10 = 80$ basis functions covering the full age-cohort surface. To constrain the β^{pq}_i values to change smoothly across age and cohort, we penalize their first-order differences. We illustrate this in Figure 3, an 8×10 grid of the basis function coefficients where the cell in row p and column q corresponds to β^{pq}

[4] Figure 2 of Camarda (2019) provides a depiction of this.

(omitting the i subscript for clarity). First-order differences can be penalized between adjacent coefficients horizontally or vertically in the cohort or age directions respectively (represented respectively by vertical or horizontal lines between cells).

We color code Figure 3 to indicate areas with different types of penalization. We constrain the coefficients in the first and last rows (rows 1 and P) to equal row-specific values β^1 and β^P respectively (gray areas):

$$\beta^{11} = \cdots = \beta^{1Q} = \beta^1; \tag{3}$$

$$\beta^{P1} = \cdots = \beta^{PQ} = \beta^P. \tag{4}$$

This allows us to achieve forecast plausibility in two ways: firstly, by borrowing strength across the youngest ages where there is large rate uncertainty, particularly for higher parities (panel b of Figure 1); and secondly, by ensuring negligible rates at the oldest ages.

For the remaining rows we penalize the first-order differences in the cohort direction (white and hatched areas) and the age direction (white area only). We implement these penalties through normal priors for the differences, specified in equations (5) and (6) respectively:

$$\beta^{pq} - \beta^{p(q-1)} \sim N\left(0, \sigma_C^2\right), p \in \{2, \ldots, P-1\}, q \in \{2, \ldots, Q\}; \tag{5}$$

$$\beta^{pq} - \beta^{(p-1)q} \sim N\left(0, \sigma_A^2\right), p \in \{3, \ldots, P-2\}, q \in \{1, \ldots, Q\}. \tag{6}$$

Standard deviation parameters σ_C and σ_A govern the degree of smoothing in the cohort and age directions respectively. We give $\sigma_C/0.01$ and $\sigma_A/0.01$ Student's t priors with 7 degrees of freedom; these informative priors express our belief that neighbouring coefficients should have similar values. When fitting the model, the priors in (5) and (6) combine to achieve smoothness across age and cohort simultaneously.

To respect the biological constraints of fertility we do not penalize the differences where the white and hatched areas meet, because experiments found that doing so allowed fertility to peak at implausibly old ages. Furthermore, we do not penalize the differences where the gray and white/hatched areas meet, because the coefficients in rows 1 and P take much

smaller values than those in the interior rows.

3.2.2 Data models

Next we specify the data models for Level-1 and Level-2 (Section 3.1). For birth order i, the Level-1 data model considers the births by age and cohort, b^i_{ac}, to follow a negative binomial distribution with mean equal to the relevant parity-specific exposure, E^{i-1}_{ac}, multiplied by the true conditional birth rate, λ^i_{ac}. The overdispersed Poisson interpretation of the negative binomial distribution enables extra variability to be accounted for, here through the dispersion parameter ϕ^i:

$$b^i_{ac} \sim \text{NegBinomial}\left(E^{i-1}_{ac}\lambda^i_{ac}, \phi^i\right).$$

Smaller values of ϕ correspond to greater overdispersion, i.e. higher variance. We specify an informative half-normal $N^+(0,5^2)$ prior for each ϕ^i to incorporate our prior belief that there is substantial uncertainty about the parity-specific rate estimates resulting from the denominator approximation (Section 2.2). In this way, the negative binomial distribution accounts for error in the measurement of the exposures, as well as sampling error.

Analogously, the Level-2 data model considers the total births by age and cohort, b_{ac}, to follow a negative binomial distribution with mean equal to the exposure by age and cohort, E_{ac}, multiplied by the true ASFR, defined as μ_{ac}:

$$b_{ac} \sim \text{NegBinomial}\left(E_{ac}\mu_{ac}, \phi^0\right).$$

We specify a vague uniform $U(0,1000^2)$ prior for the dispersion parameter ϕ^0 because unlike the parity-specific rate estimates, we can be very confident about the ASFR estimates as they only require aggregate data. Therefore large values of ϕ^0, corresponding to little overdispersion relative to the Poisson distribution, can be reasonably expected. Consequently, in contrast to the Level-1 data model, here the negative binomial distribution only accounts for sampling error.

3.3 Modelling summary fertility measures (Level-3)

In this section we specify the Level-3 system model for the summary measures, namely the true CFR and PPRs. Continuing from Section 2.3, we first consider our pooled cross-country dataset of observable CFR and PPR estimates from HFD (2024) for a given JOY value. This means that we have estimates from the 1945 cohort to the $(JOY - a_{max})$ cohort, who complete their childbearing in year JOY. Appendix E provides plots of these estimates for $JOY = 2015$. Although the country-specific curves fluctuate substantially across the cohort range, generally consecutive changes are small. We therefore base the first component of our system model on the assumption that the true CFR and PPRs change relatively slowly over time. To inform this assumption, for each summary measure SM, i.e. the CFR and different PPRs, we calculate the standard deviation of the consecutive differences from all of the country-specific curves, denoted by σ_{SM}. For $JOY = 2015$ these take the following values (to 3 significant figures):

$$\sigma_{CFR} = 0.0507;$$

$$\sigma_{PPR0\to1} = 0.0271;\ \sigma_{PPR1\to2} = 0.0301;$$

$$\sigma_{PPR2\to3} = 0.0162;\ \sigma_{PPR3+\to4+} = 0.00955.$$

To compare the standard deviations we scale them by the average value of their corresponding summary measure. We then find that actually $PPR^{2\to3}$ exhibits the greatest variability (with a scaled standard deviation of 0.0406), followed by $PPR^{1\to2}$ (0.0362), $PPR^{0\to1}$ (0.0322), $PPR^{3+\to4+}$ (0.0303) and then the CFR (0.0263).

We then express the system model for true summary measure SM, denoted by θ_c^{SM} for cohort c, as a normal prior for a first-order random walk:

$$\theta_c^{SM} \sim N\left(\theta_{c-1}^{SM}, \sigma_{SM}^2\right), c = \left(c_{min}^{JOY} + 1\right), \ldots, c_{max}^{JOY}.$$

In this way the system model indicates that values of the true summary measure for consecutive cohorts should be close to each other, with the degree of closeness informed by

the typical variation observed in the historical HFD data. The construction of this prior from observed data that already enter the model has links to Empirical Bayes methods. However, the prior will act mainly to achieve plausibility in the projected summary measures for partially observed cohorts, with any effects for fully observed cohorts likely to be negligible.

The second component of the system model relates only to $PPR^{0\to1}$, the proportion of women who have at least one child. This component is based on the assumption that very high values of the true $PPR^{0\to1}$ (and equivalently very low values of $1 - PPR^{0\to1}$, the proportion of women having no children), which indicate almost universal childbearing, are unrealistic[5]. The associated prior is informed by the pooled dataset of $PPR^{0\to1}$ estimates used above, and penalizes large values of $PPR^{0\to1}$ with increasing strength as $PPR^{0\to1}$ approaches 1. Full details are given in Appendix F.

3.4 Model fitting

For a given country and JOY value, we fit our Bayesian parity-specific fertility projection model using the package `rstan` (Stan Development Team 2024) within `R` (R Core Team 2023). This provides an interface to the `Stan` software, which enables efficient posterior sampling from complex models. We obtain 750 samples and discard the first 250 as burn-in, leaving 500 samples. Assessment of standard convergence diagnostics indicate that the samples mix well. We also fit an aggregate version that simply smooths the age-cohort ASFR surface (panel a of Figure 1) and only constrains the CFR. Both the parity-specific and aggregate models fit quickly, taking on average just over 3 minutes and less than 30 seconds respectively on a 2.7GHz Intel Core i7 Windows machine. For each posterior sample, we then generate order-specific and aggregate birth counts according to the data models in Section 3.2.2, from which we obtain posterior distributions of quantities of interest that we present in Section 4.

[5]The HFD advises caution when using estimates of $1 - PPR^{0\to1}$ which are below 5%; such estimates have been calculated for some Eastern European countries for cohorts born in the 1950s and 1960s (e.g., see Kostova et al. 2021).

3.5 Comparator methods

We assess the predictive performance of our proposed parity-specific model by comparing: (1) against the aggregate version (Section 3.4) to directly evaluate the effect of incorporating parity information, and (2) against a selection of the best-performing (aggregate) methods in the existing literature to determine whether our parity-specific model has competitive forecast accuracy and precision. We consider four comparator methods:

- Freeze rates approach: simply fixes the ASFRs at their most recent observed values.
- Model of Myrskylä et al. (2013): extrapolates the recent age-specific ASFR trend for five years and then freezes the rates.
- Model of Schmertmann et al. (2014a)[6]: a Bayesian approach which incorporates prior information from historical HFD data regarding typical ASFR patterns. While our approach performs 2D smoothing of the (parity-specific) rates explicitly, the model of Schmertmann et al. (2014a) does not smooth the observed rates, and obtains smooth forecasts implicitly by penalizing implausible future ASFR age-cohort progressions.
- model of Ševčíková et al. (2016)[7]: takes a top-down approach (see Section 1), deconstructing TFR projections from the hierarchical Bayesian model used in the United Nations population projections (Gerland et al. 2014) into ASFR projections.

The models of Myrskylä et al. (2013), Schmertmann et al. (2014a) and Ševčíková et al. (2016) were found to forecast completed cohort fertility with the highest accuracy in the recent comparative paper of Bohk-Ewald et al. (2018a). The freeze rates approach performed well despite its simplicity, leading the authors to recommend its usage as a benchmark when proposing any new fertility forecasting method.

[6] Details of the adjustments made to the implementation of this model are provided in Appendix G.
[7] Details of the adjustments made to the implementation of this model are provided in Appendix H.

4 Results

The range of model outputs presented in this section ensure that all country-JOY combinations are represented. We visualize the forecasts for eight countries which demonstrate the geographical spread and diversity of fertility patterns within our results.

4.1 Parity-specific and aggregate models

We present results from the proposed aggregate and parity-specific models fitted to Denmark, Ireland, Japan and Lithuania. We focus on $JOY = 2015$ for simplicity, with the other JOY values considered in Section 4.2 when we compare with existing methods.

4.1.1 Fitted rates

Figure 4 presents the fitted rates for Japan (rows 1 and 2) and Lithuania (rows 3 and 4). Rows 1 and 3 display the posterior median ASFRs from the two models (columns 1 and 2) followed by the parity-specific rates, which are only estimable by the parity-specific model[8]. Rows 2 and 4 plot the corresponding observed rates for comparison. The dashed lines indicate rate forecasts (row 1/3), and rate estimates observed after 2015 (row 2/4).

For Japan, 2015 follows a rapid CFR decline which stands out in a global context (Zeman et al. 2018). This is demonstrated clearly in the first two columns, where the observed ASFRs decrease steeply in the first half of the age range and rise only moderately in the second half. These columns also show reasonably similar ASFR forecasts for the aggregate and parity-specific models, with gradual declines and recuperation at younger and older ages respectively. However, the parity-specific model allows us to learn about the parity dynamics underlying these forecasts. In the subsequent columns we see that increasing smoothing is required due to the observed rates being more erratic. The further ASFR declines at younger ages appear to be driven predominantly by decreases in the rate of progression to motherhood (parity 0 rates) but also in the progression from parity 1 to 2, while all parities

[8]Note that we do not estimate parity-specific rates below age 16 for parity 1, age 17 for parity 2, and age 18 for parity 3+. This is because these rates are obsolete in the cohort cumulation approach (see Section 2.2) used to approximate the parity-specific exposures and compute the ASFRs and summary fertility measures.

(but especially parities 1 and above) contribute to the higher rates at older ages. We also note that the constraint applied for plausibility at the oldest ages (equation (4) in Section 3.2.1) ensures that the level of these increases diminishes at the upper end of the age range.

Lithuania also exhibits considerable instability in the observed ASFRs, with the ASFR collapse across ages 15–25 that occurred for the older half of the cohorts following the end of the communist regime (Billingsley 2010). Unlike Japan, however, this was counteracted by a steep rise in the ASFRs at later ages, which both models predict to continue for the younger cohorts. Interestingly, both models also anticipate a slight decrease in the level of the peak ASFR for the youngest cohorts, occurring at around age 29. The breakdown by parity determines the main drivers of these changes to be again at parities 0 and 1, with the higher-order projections remaining reasonably consistent.

Appendix I provides analogous figures for Denmark and Ireland. Denmark presents much more stable observed ASFRs, with a gentle postponement of childbearing to older ages; they are projected to remain constant at younger ages but continue to increase gradually at older ages. Also, the aggregate and parity-specific ASFR forecasts differ noticeably, with a steeper predicted rise at central ages for the former. The observed ASFRs for Ireland depict a fast but smooth fertility decline and postponement driven by steep decreases for parities 1 and above, particularly towards younger ages. Both models forecast further ASFR reductions at younger ages and rises at older ages, the aggregate model again appearing more extreme. The PPR constraints encourage more realistic parity-specific forecasts while maintaining the CFR projection, driving the more stable ASFR forecasts; this also occurs for the other countries, but less adjustment is required as these forecasts are already reasonable.

4.1.2 Summary measures

Figure 5 presents the forecasts of the five summary measures, namely the CFR and PPRs, with their associated uncertainty. In row 1 we plot the posterior median CFR and corresponding 95% prediction interval under the aggregate and parity-specific models for the four countries. We first observe that the posterior medians have reasonably similar trends, but differ in level. For Japan and Lithuania both models predict an increasing trend

followed by a plateau, with the parity-specific model rising to a slightly higher level. For Denmark and Ireland the projections diverge more considerably, the parity-specific model predicting for Denmark a decline for the late-1970s cohorts that the aggregate model does not anticipate, and for Ireland a more extreme decline for the later cohorts.

Regarding the level of uncertainty, while the intervals for the two models overlap extensively, those for the parity-specific model are consistently narrower, meaning that the projections have greater precision. Lastly, although the additional data points from 2016 onwards can be used to assess forecast performance, these correspond to cohorts who were already nearing the end of their reproductive lives and thus their CFR values can be predicted with high accuracy. In Section 4.2 we assess the forecast performance for the earlier JOY values, where longer time series of out-of-sample data are available.

We now consider the posterior distributions of the four PPRs, which are presented analogously in row 2 of Figure 5 (for the parity-specific model only). We first consider Ireland and Japan, which both exhibit a steep observed CFR decline for the complete cohorts, but due to very different trends in the observed PPRs. While Japan has seen steep declines in the progression to motherhood ($PPR^{0\to1}$) and from one to two children ($PPR^{1\to2}$), the higher-order PPRs have decreased fastest for Ireland. This supports our earlier observations that the greatest change in the parity-specific rates was exhibited for parities 0 and 1 in Japan, but parities 1 and above in Ireland. Concerning the projections, while the higher-order PPRs are predicted to decline further in Ireland and so continue to reduce the CFR with $PPR^{0\to1}$ remaining reasonably stable, the lower-order PPRs (and the CFR) are projected to recover and somewhat stabilize for Japan.

Denmark displays a gradual CFR increase for the complete cohorts, which appears to be driven by similar observed rises in $PPR^{0\to1}$ and $PPR^{1\to2}$, the higher-order PPRs remaining stable. However declines are initially predicted for all PPRs, with recovery anticipated for the lower-order PPRs while the higher-order PPRs plateau. Lithuania also presents relatively stable observed PPR trends, despite great instability in the underlying rates (Section 4.1.1). Again, recovery is projected for the lower-order PPRs, while the higher-order PPRs are

predicted to converge. The steep projected increase in $PPR^{1\to 2}$ drives the rise in the CFR, leading to the predicted CFR convergence for the youngest cohorts in Denmark and Lithuania. In general, the gradual nature of these predicted PPR and CFR changes reflects the addition of their respective constraints to incorporate prior knowledge regarding their relatively slow progressions in the past (Section 3.3).

4.2 Comparison with existing methods

In continuation from Section 3.5, in this subsection we assess the predictive performance of our proposed parity-specific model by comparing with the aggregate model and the four comparator methods from the existing literature.

4.2.1 Summary statistics

We first compare the predictive performance of these methods across all countries and *JOY* values simultaneously using five different summary statistics relating to the ASFR and CFR forecasts, which we present in Table 1. Note that of the four comparator methods from the existing literature, only the models of Schmertmann et al. (2014a) and Ševčíková et al. (2016) are taken to have probabilistic forecasts[9]. The first three statistics are standard measures, namely the mean absolute error (MAE), mean absolute percentage error (MAPE) and the root mean square error (RMSE), where we take the point estimate for the probabilistic forecasts to be the posterior median of the forecast distribution[10]. For both the ASFR and CFR forecasts, these measures indicate that our proposed parity-specific model is preferred marginally over the aggregate version; for the ASFR forecasts the model of Ševčíková et al. (2016) performs slightly better than the parity-specific model, but for the CFR forecasts the four comparator methods perform worse, in some cases quite considerably.

The next two rows contain the coverage of the 90% and 50% prediction intervals for the four

[9] Although Myrskylä et al. (2013) provide a method for computing forecast intervals, their coverage has been shown to be poor (Bohk-Ewald et al. 2018a; Ellison et al. 2020). We therefore only consider the point forecasts of this approach.

[10] Note that to ensure a fair comparison between the deterministic models (which provide a perfect fit to the observed ASFRs) and the probabilistic models, the CFR summary statistics for the latter in Table 1 (excluding the coverages) and Figure 6 ignore the uncertainty about the ASFRs observed before the jump-off year, instead treating them as known.

probabilistic methods. The values for the ASFR forecasts are considerably lower than the nominal levels, with the Ševčíková et al. (2016) model, and the aggregate and Schmertmann et al. (2014a) models (jointly) having the highest coverage for the 90% and 50% prediction intervals respectively. Contrastingly, the values for the CFR forecasts are generally much closer to the nominal levels, with the proposed parity-specific and aggregate models well calibrated at the 90% level but with slight overcoverage at the 50% level.

The final row presents the average value of a particular scoring rule. Scoring rules have been applied in the context of fertility forecasting by Ellison et al. (2020). They are designed to assess probabilistic rather than point forecasts and so can measure the performance of the entire forecast distribution (Gneiting and Raftery 2007). We compute the continuous ranked probability score (CRPS) (Matheson and Winkler 1976), which evaluates the quality of a forecast distribution based on how close its cumulative distribution function (CDF) is to that of the true value; it can also be calculated where only point forecasts are available (and is equivalent to the MAE), meaning that we can compare all of our methods simultaneously (Ellison et al. 2020). We note that smaller CRPS values indicate greater predictive accuracy. Table 1 indicates that the ASFR forecasts from the models of Schmertmann et al. (2014a) and Myrskylä et al. (2013) are most preferable under the CRPS, with our parity-specific model reasonably close. However, for the CFR forecasts the parity-specific model is preferable.

The summary statistics imply that overall, our proposed parity-specific model tends to outperform the corresponding aggregate model and is also competitive with some of the best-performing existing models. However this is at a high level of aggregation, and so next we summarize the differences we observe when repeating this exercise for a given country or JOY value (results not shown). The conclusion regarding the comparison with the aggregate model generally holds, with only a few instances where the aggregate model is preferable based on at least one measure (ASFR: Estonia, Japan, Norway and Taiwan; CFR: Denmark, Ireland, Japan, the Netherlands and the United States, and 2005). The findings regarding the comparison with the existing methods are less consistent across countries and JOY values. Across the ASFR forecasts, the model of Schmertmann et al. (2014a) is generally the most preferable, with the model of Myrskylä et al. (2013) only preferable for the United States and

2015; for the CFR forecasts, the parity-specific model is strongly preferable.

Next, in Figure 6 we inspect the results broken down for each country-JOY combination for the CFR forecasts. In panel a we plot the average CRPS by country for all JOY values combined, while panel b provides plots for each individual JOY value. The countries are ordered by decreasing CRPS according to the parity-specific model, to allow ease of comparison relative to the alternative methods. We first contrast the lines corresponding to the parity-specific and aggregate models, observing that in the vast majority of cases where there is a noticeable difference, the parity-specific model has the lower average CRPS; this supports our findings from Table 1.

Secondly, we contrast the parity-specific model with the remaining lines, corresponding to the four comparator methods. Considering the 'benchmark' freeze rates approach first, our parity-specific model strongly outperforms this method overall for nearly all countries (panel a) and for JOY values 2000 and 2005 (panel b); the methods are more evenly matched for the later JOY values. This is intuitive because the averages are dominated by completing cohort fertility at high truncation ages, which is reasonably straightforward to achieve with considerable accuracy; this explains the more erratic results and greater crossover between the methods. For the models of Ševčíková et al. (2016) and Schmertmann et al. (2014a) the parity-specific model is still preferable in general, while the model of Myrskylä et al. (2013) often outperforms the parity-specific model.

4.2.2 Visual assessment

Next we visually inspect the projections, starting with plotting the CFR forecast distributions for our second selection of four countries – Canada, Czechia, the Netherlands and the United States – for JOY values 2000 and 2005 in rows 1 and 2 of Figure 7. We choose these earliest jump-off years as they provide the longest time series of out-of-sample data points. In most cases the prediction intervals or point forecasts for the five methods overlap somewhat. However, the 2000-based forecasts for Czechia provide an interesting example whereby the predictions diverge sharply from around the 1970 cohort; the parity-specific and aggregate approaches anticipate a recovery to replacement level, while the other

methods predict a steep decline to unprecedented levels. The reality lies in the middle, with a gradual CFR decline being observed that lies much closer to the point forecasts from our proposed models; this explains the poor performance of the other methods in Figure 6.

In contrast, the 2000-based forecasts for the United States demonstrate an instance where the simple extrapolative method of Myrskylä et al. (2013) performs best; the parity-specific, aggregate and Ševčíková et al. (2016) models predict a fairly similar trend, while the freeze rates approach and the model of Schmertmann et al. (2014a) respectively understate and overstate the the extent of the CFR recovery. This latter model underperforms in a similar way for the 2000-based Netherlands and 2005-based United States forecasts; however, it outperforms the other methods in the 2000-based Canada and 2005-based Czechia forecasts. This is consistent with the model exhibiting some of the widest ranging average CRPS values in Figure 6.

Lastly, in rows 3–5 of Figure 7 we examine the ASFR forecasts for the 1975 cohort at *JOY* values 2000, 2005 and 2010, when women in this cohort were aged 25, 30 and 35 respectively. This allows us to appreciate the challenge of completing the cohort schedule at young truncation ages. For example, the cause of the divergent 2000-based Czechia CFR forecasts (row 1 of Figure 7) is clear looking at the underlying ASFRs for the 1975 cohort – the age pattern has peaked and is stable, but when it will decline (or whether it will peak further) is difficult to determine, hence the very different forecasts. However, the projections for the same cohort at the later *JOY* values exhibit much greater consistency due to a larger proportion of the cohort schedule having been observed. A similar pattern is visible in the Netherlands panels, with the methods disagreeing on the location and height of the ASFR peak in the 2000-based forecast, but mostly agreeing in the later forecasts. The same is true for Canada and the United States but to a lesser extent.

5 Discussion and conclusions

In this paper we propose, to the best of our knowledge, the first integrated methodological framework for producing fertility projections by parity. The model generates smooth

age-cohort surfaces of rates for each parity which are combined to give an aggregate rate surface, allowing summary fertility measures such as completed family size (via the cohort total fertility rate, or CFR) and parity progression ratios to be computed within the model. Taking a Bayesian approach makes it straightforward to apply constraints both to these measures and to the smooth surfaces themselves, thus incorporating prior knowledge and improving forecast plausibility; it also allows the model to be informed by data sources at the parity-specific and aggregate levels in a coherent way.

The key methodological challenge to overcome is how to model *conditional* parity-specific fertility rates, i.e. rates that depend on the number of women at risk of each particular birth order, when in most cases data of sufficient quality are only available for the numerator, i.e. the number of order-specific births. To this end, we exploit the sequential progression of parity through a person's life and approximate the denominators by cumulating across cohort fertility (Jasilioniene et al. 2015; Smallwood 2002). This is similar to the approach taken by Ellison et al. (2023), however an important difference is that our cumulation occurs during model fitting rather than only being applied post-hoc to separate parity-specific model outputs. The implementation of such an approximation is necessary even if the denominators for the fitting period are known, as estimates consistent with the rate projections are required for the forecast period. Despite this, the fact remains that fitting the proposed model relies on having at least some parity-specific data available at the population level, which may pose a difficulty in less-developed countries.

A second challenge is how to obtain forecasts that are plausible, especially given that we are extrapolating multiple two-dimensional smooth surfaces within the same model. We do this by applying various constraints that express our prior knowledge about the likely progression of demographic rates (see Camarda 2019 for an example of such an approach in the context of mortality forecasting). The first instance is at the parity-specific level (Section 3.2.1), constraining the parameters to borrow strength across cohorts where the rates are erratic at the youngest ages, and to enforce the known biological limitations of fertility at the oldest ages. The second application is at the aggregate level (Section 3.3), where we (1) constrain summary fertility measures to change slowly from cohort to cohort,

informed by the historical variability of these measures in a pooled dataset from the Human Fertility Database (HFD) (HFD 2024); and (2) discourage unrealistically high values of the proportion of women becoming mothers. These constraints, when taken together, should lead to increased forecast plausibility in terms of the shapes of the cohort schedules of rates at the parity-specific level, and the progression of summary measures at the aggregate level.

To assess the forecast performance of our proposed model we fit it to 16 countries across four cohort ranges with a different jump-off year, i.e. the last year of data used for model fitting. We also fit an analogous 'aggregate' model, which only smooths the overall age-specific fertility rates (ASFRs) and constrains the CFR, as well as several competitive aggregate models from the literature, informed by the results of Bohk-Ewald et al. (2018a). We perform a comprehensive quantitative and visual comparison of the resulting ASFR and CFR forecasts to ascertain whether the incorporation of parity information improves forecast accuracy and precision. We can address this question directly by comparing the parity-specific and aggregate models, which only differ in this respect; we find that the former is strongly preferable under nearly all of the measures. In contrast, the comparison with existing approaches has mixed findings, with the comparator methods often being preferable, particularly for the ASFR forecasts. However, this comparison only addresses the question indirectly – there are many differences between our proposed model and the existing methods which do not relate to parity and cannot be eliminated.

We have tested the sensitivity of our proposed model in several ways, considering a more flexible prior on the summary measures (Appendix J), different knot spacing combinations for the age and cohort basis functions (Appendix K), and the use of register- or census-based age-parity population estimates rather than the aforementioned cumulation-based approach (Appendix L). In each case we find that no alternative implementation provides a consistent improvement in predictive accuracy and coverage across the countries and jump-off years considered. These preliminary investigations demonstrate that our proposed model appears to be reasonably robust to the choice of exposure data, smoothing and prior specifications. Further refinement could include testing the sensitivity to the other constraints and priors within the model.

Although we include as many HFD countries as possible, it is necessary to exclude those with lower-quality order-specific births data. Furthermore, the HFD only includes countries with high-quality birth and population data, which tends to be highly developed countries (Jasilioniene et al. 2016). Therefore, a key area for future investigation is to apply the model to countries with poorer-quality or absent order-specific birth data. Our proposed Bayesian framework makes it reasonably straightforward to handle the various issues that could arise. For example, coarser order-specific births data can be incorporated by aggregating the relevant modelled quantities, adjusting the data models to reflect the level of reliability. Higher-quality data from similar countries can be integrated to borrow strength, for instance by specifying priors regarding their typical parity-specific age pattern. This demonstrates the possibility of obtaining plausible estimates and projections in data-poor contexts by adapting our proposed model to incorporate non-standard data, different levels of confidence and additional data sources.

Further model extensions could include the incorporation of time since last birth to capture the important dependence of birth events on this variable (Ellison et al. 2022), and the adaptation to a period context. Furthermore, scenario testing could be of great interest to policymakers, whereby the effect of various parity-specific interventions on overall childbearing levels could be explored. Lastly, our proposed model could be incorporated into a kinship projection model, where advancements have already been made through an efficient matrix formulation and the inclusion of parity as a characteristic (Caswell 2020).

In conclusion, our proposed model provides a sophisticated and efficient Bayesian methodology for obtaining parity-specific fertility projections. That such projections can express aggregate rates via the underlying parity dynamics that interact to produce them undoubtedly adds great richness and incorporates an awareness of the sequential nature of childbearing, features which are absent from existing methods. Furthermore, the ability to generate plausible projections of parity-specific summary measures, including parity progression ratios and family size distributions, is exciting in and of itself. Of equal importance is forecast performance, which we find to be reasonable when compared to some of the best-performing models from the literature. This demonstrates that enhancing

explanatory capacity, in this case through explicitly modelling the childbearing mechanism by parity, yields a model that is at least comparable in terms of its predictive performance and errors. This provides a strong argument for the use and further development of fertility projections that more accurately capture the process of childbearing, thus having the potential to enable more tailored policy solutions.

Acknowledgements

This work was partly supported by the Economic and Social Research Council (ESRC) FertilityTrends project (grant ES/S009477/1) and the Centre for Population Change Connecting Generations Centre (grant ES/W002116/1). Earlier versions of this work have been presented at the Nordic Demographic Symposium (June 2022), the European Population Conference (July 2022), the British Society for Population Studies conference and the Royal Statistical Society Annual Conference (both September 2022). The authors thank Ann Berrington, Sarah Christison, Bernice Kuang, Hill Kulu, Jon Forster, Jason Hilton and Peter W. F. Smith for their feedback during the research phase of the project. The authors are also grateful to the editor and four anonymous reviewers for their helpful comments on an earlier version of this paper.

References

Aburto, J. M., Kashyap, R., Schöley, J., Angus, C., Ermisch, J., Mills, M. C., and Dowd, J. B. (2021). Estimating the burden of the COVID-19 pandemic on mortality, life expectancy and lifespan inequality in England and Wales: a population-level analysis. *Journal of Epidemiology and Community Health*, 75(8):735–740.

Ahlburg, D. A. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*, chapter Population Forecasting, pages 557–575. Springer, Boston, MA.

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T., and Heilig, G. K. (2011). Probabilistic Projections of the Total Fertility Rate for All Countries. *Demography*, 48:815–839.

Andersson, G. (1999). Childbearing Trends in Sweden 1961–1997. *European Journal of Population*, 15:1–24.

Andreev, E. M., Shkolnikov, V. M., and Begun, A. Z. (2002). Algorithm for decomposition of differences between aggregate demographic measures and its application to life expectancies, healthy life expectancies, parity-progression ratios and total fertility rates. *Demographic Research*, 7:499–522.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., and Panagiotelis, A. (2024). Forecast reconciliation: A review. *International Journal of Forecasting*, 40(2):430–456.

Begall, K. and Mills, M. C. (2013). The Influence of Educational Field, Occupation, and Occupational Sex Segregation on Fertility in Netherlands. *European Sociological Review*, 29:720–742.

Billingsley, S. (2010). The Post-communist Fertility Puzzle. *Population Research and Policy Review*, 29(2):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(37):9187–9192.

Bohk-Ewald, C., Li, P., and Myrskylä, M. (2018b). *Forecast accuracy hardly improves with method complexity when completing cohort fertility* [R code]. https://github.com/fertility-forecasting/validate-forecast-methods. Accessed 4 January 2019.

Bongaarts, J. and Sobotka, T. (2012). A Demographic Explanation for the Recent Rise in European Fertility. *Population and Development Review*, 38(1):83–120.

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 22:547–581.

Camarda, C. G. (2019). Smooth constrained mortality forecasting. *Demographic Research*, 41(38):1091–1130.

Caswell, H. (2020). The formal demography of kinship II: Multistate models, parity, and sibship. *Demographic Research*, 42(38):1097–1146.

Currie, I. D., Durban, M., and Eilers, P. H. C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4:279–298.

de Beer, J. (1985). A time series model for cohort data. *Journal of the American Statistical Association*, 80(391):525–530.

Duarte, C. and Rua, A. (2007). Forecasting inflation through a bottom-up approach: How bottom is bottom? *Economic Modelling*, 24(6):941–953.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2):89–121.

Ellison, J., Berrington, A., Dodd, E., and Forster, J. J. (2022). Investigating the application of generalized additive models to discrete-time event history analysis for birth events. *Demographic Research*, 47:647–694.

Ellison, J., Berrington, A., Dodd, E., and Forster, J. J. (2023). Combining individual- and population-level data to develop a Bayesian parity-specific fertility projection model. *Journal of the Royal Statistical Society: Series C*, qlad095.

Ellison, J., Dodd, E., and Forster, J. J. (2020). Forecasting of cohort fertility under a hierarchical Bayesian approach. *Journal of the Royal Statistical Society: Series A*, 183(3):829–856.

Feeney, G. and Yu, J. (1987). Period Parity Progression Measures of Fertility in China. *Population Studies*, 41(1):77–102.

Gerland, P., Raftery, A. E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B. K., Chunn, J., Lalic, N., Bay, G., Buettner, T., Heilig, G. K., and Wilmoth, J. (2014). World population stabilization unlikely this century. *Science*, 346:234–237.

Gleditsch, R. F., Syse, A., and Thomas, M. J. (2021). Fertility Projections in a European

Context: A Survey of Current Practices among Statistical Agencies. *Journal of Official Statistics*, 37(3):547–568.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Greulich, A. and Toulemon, L. (2023). Measuring the educational gradient of period fertility in 28 European countries. *Demographic Research*, 49:905–968.

Heinisch, K. and Scheufele, R. (2018). Bottom-up or direct? Forecasting German GDP in a data-rich environment. *Empirical Economics*, 54:705–745.

Hellstrand, J., Nisén, J., Miranda, V., Fallesen, P., Dommermuth, L., and Myrskylä, M. (2021). Not Just Later, but Fewer: Novel Trends in Cohort Fertility in the Nordic Countries. *Demography*, 58(4):1373–1399.

Hellstrand, J., Nisén, J., and Myrskylä, M. (2020). All-time low period fertility in Finland: Demographic drivers, tempo effects, and cohort implications. *Population Studies*, 74(3):315–329.

Hilton, J., Dodd, E., Forster, J. J., and Smith, P. W. F. (2019). Projecting UK mortality by using Bayesian generalized additive models. *Journal of the Royal Statistical Society: Series C*, 68:29–49.

Hilton, J., Dodd, E., Forster, J. J., and Smith, P. W. F. (2021). Modelling Frontier Mortality Using Bayesian Generalised Additive Models. *Journal of Official Statistics*, 37(3):569–589.

Hobcraft, J. (1996). Fertility in England and Wales: A Fifty-Year Perspective. *Population Studies*, 50(3):485–524.

Human Fertility Database (HFD) (2024). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). http://www.humanfertility.org. Data downloaded 2 October 2024.

Jasilioniene, A., Jdanov, D. A., Sobotka, T., Andreev, E. M., Zeman, K., Shkolnikov, V. M., and with contributions from J Goldstein, E J Nash, D Philipov, and G Rodriguez (2015).

Methods Protocol for the Human Fertility Database. https://www.humanfertility.org/File/GetDocumentFree/Docs/methods.pdf. Accessed 6 November 2023.

Jasilioniene, A., Sobotka, T., Jdanov, D. A., Zeman, K., Kostova, D., Andreev, E. M., Grigoriev, P., and Shkolnikov, V. M. (2016). Data Resource Profile: The Human Fertility Database. *International Journal of Epidemiology*, 45(4):1077–1078e.

Kohler, H.-P. and Ortega, J. A. (2002). Tempo-adjusted period parity progression measures, fertility postponement and completed cohort fertility. *Demographic Research*, 6:91–144.

Kostova, D., Grigoriev, P., and revised by Jasilioniene, A (2021). *Human Fertility Database Documentation: Belarus*. https://www.humanfertility.org/File/GetDocumentFree/Docs/BLR/BLRcom.pdf. Accessed 14 March 2025.

Kravdal, Ø. (2001). The High Fertility of College Educated Women in Norway: An Artefact of the Separate Modelling of Each Parity Transition. *Demographic Research*, 5:187–216.

Li, H. and Hyndman, R. J. (2021). Assessing mortality inequality in the U.S.: What can be said about the future? *Insurance: Mathematics and Economics*, 99:152–162.

Li, H., Li, H., Lu, Y., and Panagiotelis, A. (2019). A forecast reconciliation approach to cause-of-death mortality modeling. *Insurance: Mathematics and Economics*, 86:122–133.

Li, N. and Wu, Z. (2003). Forecasting cohort incomplete fertility: A method and an application. *Population Studies*, 57(3):303–320.

Liu, P. and Raftery, A. E. (2020). Accounting for uncertainty about past values in probabilistic projections of the total fertility rate for most countries. *The Annals of Applied Statistics*, 14(2):685–705.

Lutz, W., Goujon, A., and Doblhammer-Reiter, G. (1998). Demographic dimensions in forecasting: Adding education to age and sex. *Population and Development Review*, 24:42–58.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096.

Mazzuco, S. and Keilman, N. (2020). *Developments in Demographic Forecasting*. Springer Nature.

McDonald, P., Hosseini-Chavoshi, M., Abbasi-Shavazi, M. J., and Rashidian, A. (2015). An assessment of recent Iranian fertility trends using parity progression ratios. *Demographic Research*, 32:1581–1602.

Murphy, M. and Berrington, A. (1993). Constructing period parity progression ratios from household survey data. In Ní Bhrolcháin, M., editor, *New Perspectives on Fertility in Britain*, pages 17–33. London: HMSO.

Myrskylä, M., Goldstein, J. R., and Cheng, Y. A. (2013). New cohort fertility forecasts for the developed world: Rises, falls, and reversals. *Population and Development Review*, 39(1):31–56.

Ní Bhrolcháin, M. (1987). Period Parity Progression Ratios and Birth Intervals in England and Wales, 1941–1971: A Synthetic Life Table Analysis. *Population Studies*, 41(1):103–125.

Ní Bhrolcháin, M. (1992). Period paramount? A critique of the cohort approach to fertility. *Population and Development Review*, 18(4):599–629.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, 38(3):705–871.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Lewis, S. M., Aghajanian, A., and Kahn, M. J. (1996). Event history modeling of world fertility survey data. *Mathematical Population Studies*, 6(2):129–153.

Rallu, J.-L. and Toulemon, L. (1994). Period Fertility Measures: The Construction of

Different Indices and their Application to France, 1946-89. *Population: An English Selection*, 6:59–93.

Schmertmann, C., Zagheni, E., Goldstein, J. R., and Myrskylä, M. (2014a). Bayesian Forecasting of Cohort Fertility. *Journal of the American Statistical Association*, 109(506):500–513.

Schmertmann, C., Zagheni, E., Goldstein, J. R., and Myrskylä, M. (2014b). *Bayesian Forecasting of Cohort Fertility* [project website]. http://schmert.net/cohort-fertility/. Accessed 1 November 2023.

Schoen, R. (2022). Relating period and cohort fertility. *Demography*, 59(3):877–894.

Ševčíková, H., Alkema, L., and Raftery, A. E. (2011). `bayesTFR`: An R Package for Probabilistic Projections of the Total Fertility Rate. *Journal of Statistical Software*, 43:1–29.

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

Ševčíková, H. and Raftery, A. E. (2016). `bayesPop`: Probabilistic Population Projections. *Journal of Statistical Software*, 75:1–29.

Shang, H. L. and Booth, H. (2020). Synergy in fertility forecasting: improving forecast accuracy through model averaging. *Genus*, 76(27).

Smallwood, S. (2002). New estimates of trends in births by birth order in England and Wales. *Population Trends*, 108:32–48.

Sobotka, T. (2003). Tempo-quantum and period-cohort interplay in fertility changes in Europe. Evidence from the Czech Republic, Italy, the Netherlands and Sweden. *Demographic Research*, 8:151–214.

Stan Development Team (2024). *`RStan`: The `R` interface to `Stan`*. R package version 2.32.5. http://mc-stan.org/.

United Nations, Department of Economic and Social Affairs, Population Division (2024). *World Population Prospects 2024, Online Edition*. https://population.un.org/wpp/. Data downloaded 5 November 2024.

Van Hook, J. and Altman, C. E. (2013). Using Discrete-Time Event History Fertility Models to Simulate Total Fertility Rates and Other Fertility Measures. *Population Research and Policy Review*, 32:585–610.

van Raalte, A. A., Basellini, U., Camarda, C. G., Nepomuceno, M. R., and Myrskylä, M. (2023). The dangers of drawing cohort profiles from period data: A research note. *Demography*, 60(6):1689–1698.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall/CRC Press.

Zeman, K., Beaujouan, E., Brzozowska, Z., and Sobotka, T. (2018). Cohort fertility decline in low fertility countries: Decomposition using parity progression ratios. *Demographic Research*, 38:651–690.

Zotteri, G., Kalchschmidt, M., and Caniato, F. (2005). The impact of aggregation level on forecasting performance. *International Journal of Production Economics*, 93–94:479–491.

Table 1: Summary statistics for the ASFR and CFR forecasts under various models

Forecast	Measure	P-S	A	Šev	Sch	M	FR
ASFR	MAE	0.0063	0.0073	**0.0058**	0.0078	0.0067	0.0073
	MAPE (%, 1 decimal place)	**11.7**	13.5	12.2	18.3	13.3	16.5
	RMSE	0.010	0.011	**0.009**	0.012	0.011	0.011
	Coverage of 90% PI (%)	60	68	**71**	66	-	-
	Coverage of 50% PI (%)	28	**35**	34	**35**	-	-
	Average CRPS	0.0071	0.0086	0.0071	**0.0060**	0.0067	0.0073
CFR	MAE	**0.012**	0.013	0.019	0.026	0.015	0.028
	MAPE (%, 1 decimal place)	**0.7**	**0.7**	1.0	1.4	0.8	1.5
	RMSE	**0.029**	0.030	0.041	0.056	0.038	0.060
	Coverage of 90% PI (%)	**87**	94	70	71	-	-
	Coverage of 50% PI (%)	61	67	36	**51**	-	-
	Average CRPS	**0.013**	0.015	0.022	0.022	0.015	0.028

Notes: Summary statistics are calculated across all countries and *JOY* values for the ASFR and CFR forecasts under the proposed parity-specific (P-S) and aggregate (A) models, the Ševčíková et al. (Šev), Schmertmann et al. (Sch) and Myrskylä et al. (M) models, and the freeze rates approach (FR). Values are given to 2 significant figures unless stated otherwise. MAE = mean absolute error; MAPE = mean absolute percentage error; RMSE = root mean square error; PI = prediction interval; CRPS = continuous ranked probability score. The most preferable value/values is/are in **bold**.

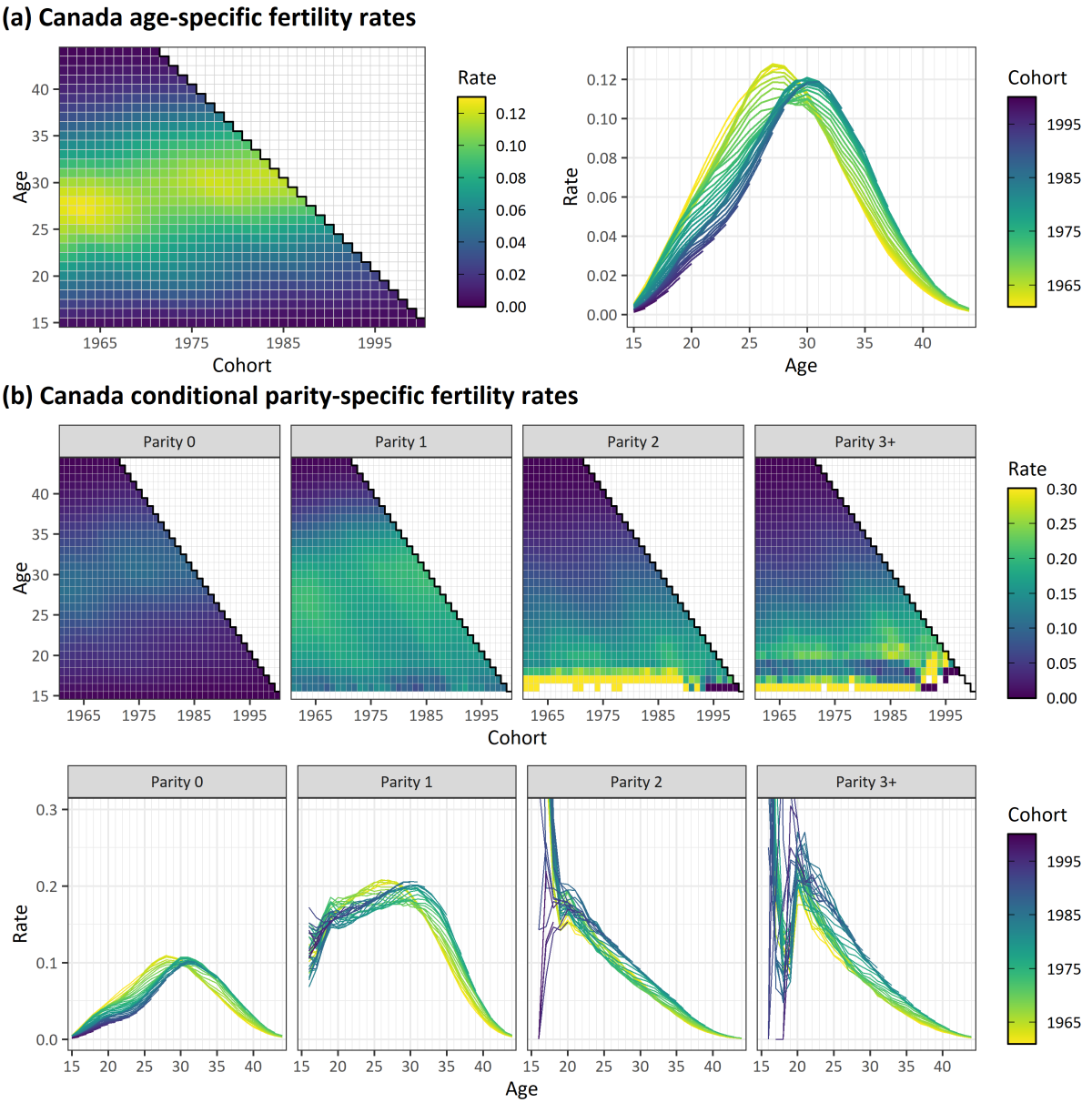

Figure 1: (a) Left: Lexis surface of Canada age-specific fertility rate estimates for the 1961–2000 cohorts, observed up to $JOY = 2015$, from HFD (2024). Right: same data presented in the form of cohort-specific rate curves. (b) Top row: Lexis surfaces of Canada conditional parity-specific fertility rate estimates for the 1961–2000 cohorts and parities 0, 1, 2 and 3+, observed up to $JOY = 2015$, from HFD (2024); note that the rate estimates have been top-coded at 0.3. Bottom row: same data presented in the form of cohort-specific rate curves. Note the non-standard age-cohort axes in the Lexis surfaces.

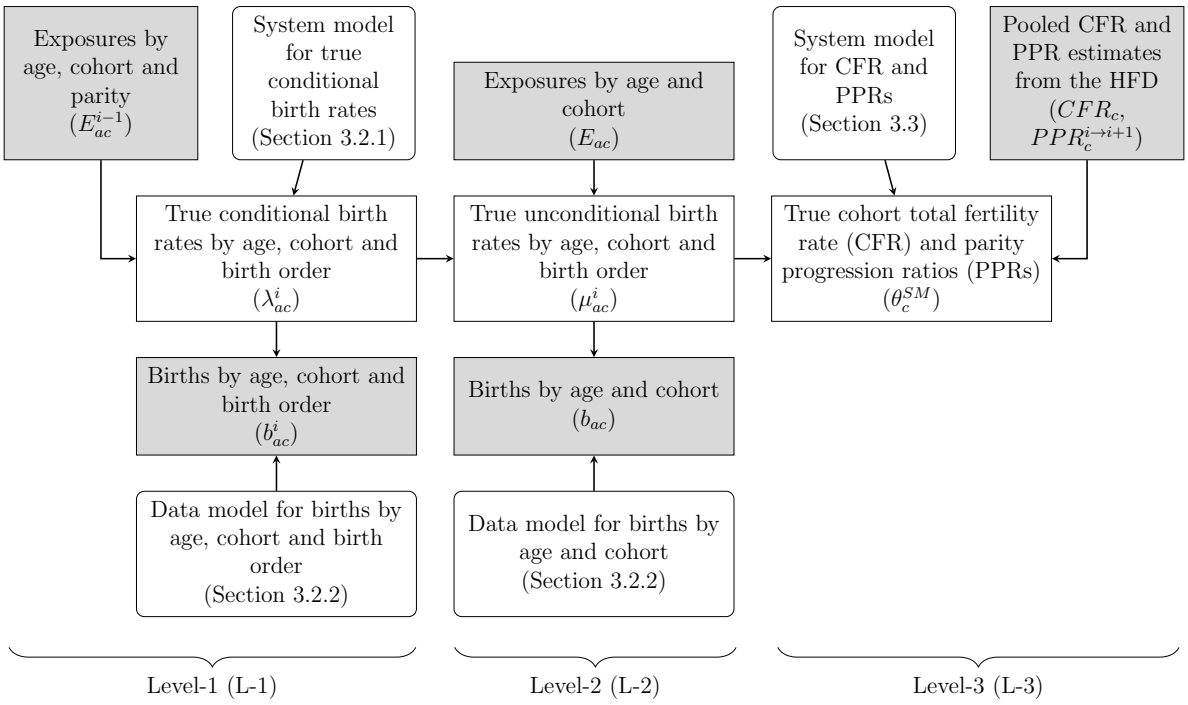

Figure 2: Overview of the proposed modelling framework. Straight-edged rectangles are demographic arrays, rounded rectangles are system and data models; only the gray rectangles are observed. For convenience, the corresponding symbol (straight-edged rectangles) and section number where the corresponding system or data model is specified (rounded rectangles) are given in parentheses.

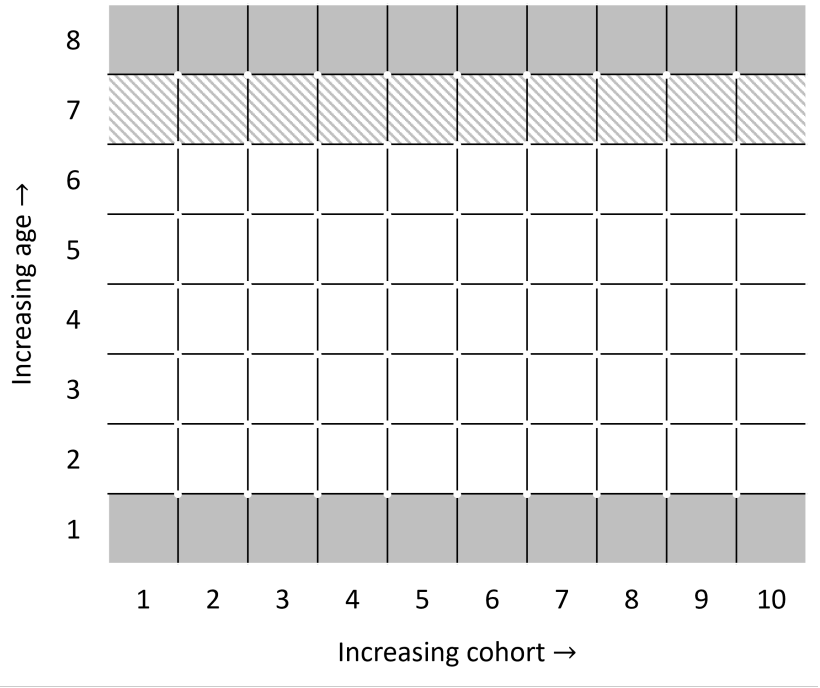

Figure 3: Illustration of penalties applied to basis function coefficients for a 2D age-cohort surface with $P = 8$ basis functions across age and $Q = 10$ across cohort. Gray areas: coefficients fixed across cohort; White and hatched areas: first-order differences penalized in cohort direction; White area: first-order differences penalized in age direction.

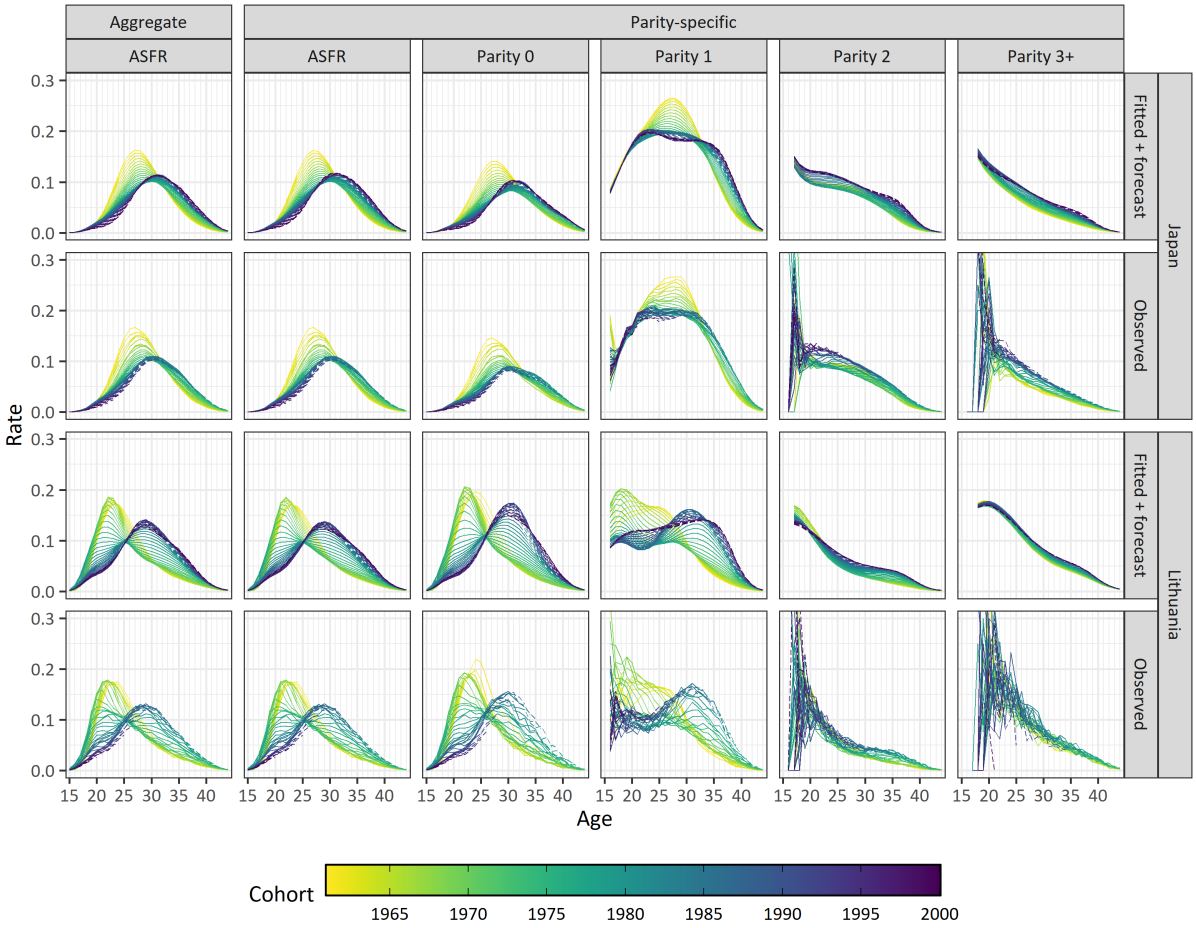

Figure 4: Row 1: Japan posterior median ASFRs and parity-specific rates from the aggregate and parity-specific fertility projection models for $JOY = 2015$; Row 2: corresponding observed rates from HFD (2024). Rows 3 and 4: Equivalent plots for Lithuania. Dashed lines indicate values after the year JOY.

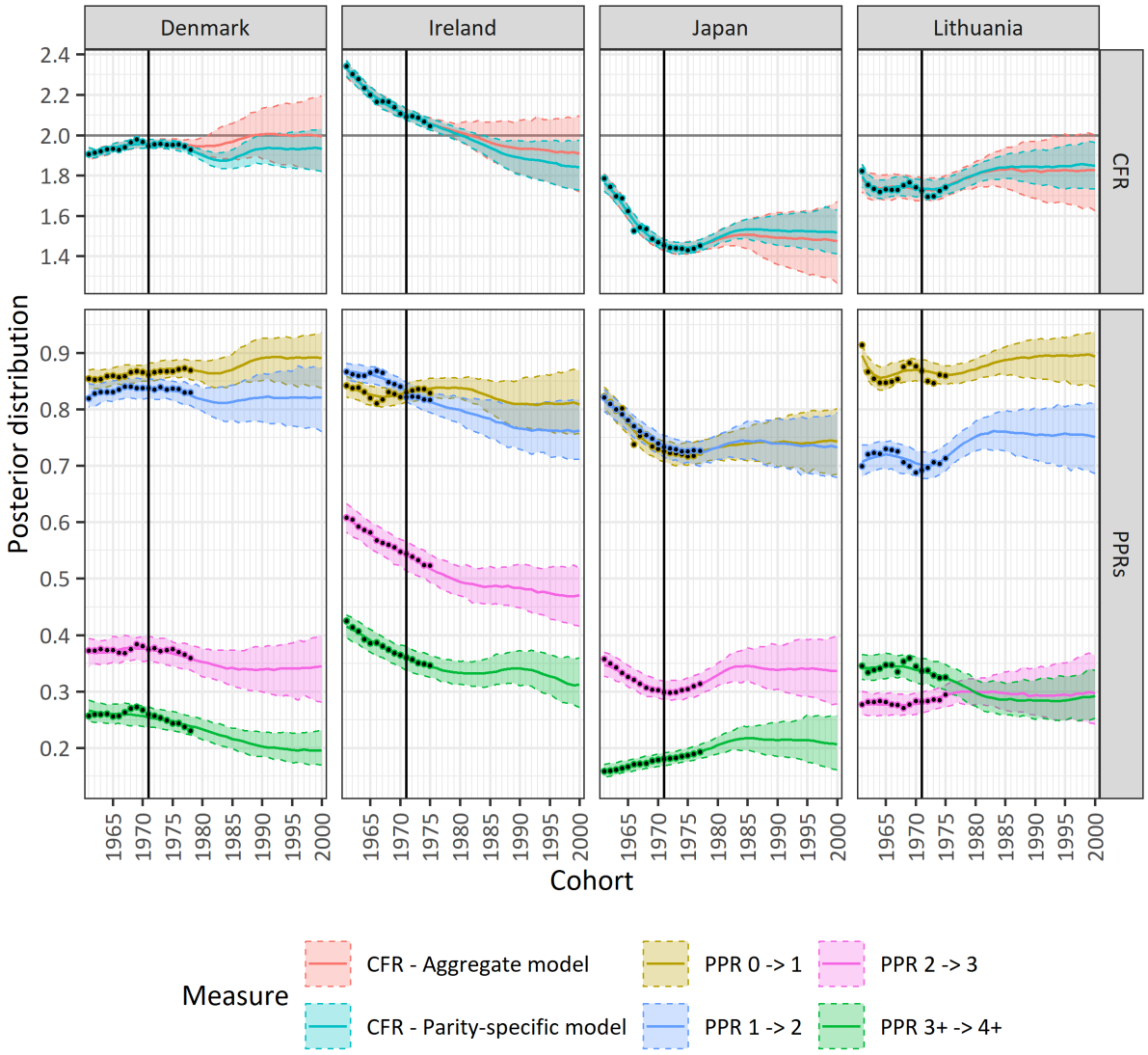

Figure 5: CFR (top row) and PPR (bottom row) posterior medians and 95% prediction intervals for the aggregate and parity-specific fertility projection models for $JOY = 2015$ and four selected countries; note that the PPRs are only estimable by the parity-specific model. Points are observed values from HFD (2024); vertical lines indicate the last fully-observed cohort (born in the year 1971 = $JOY - a_{max}$).

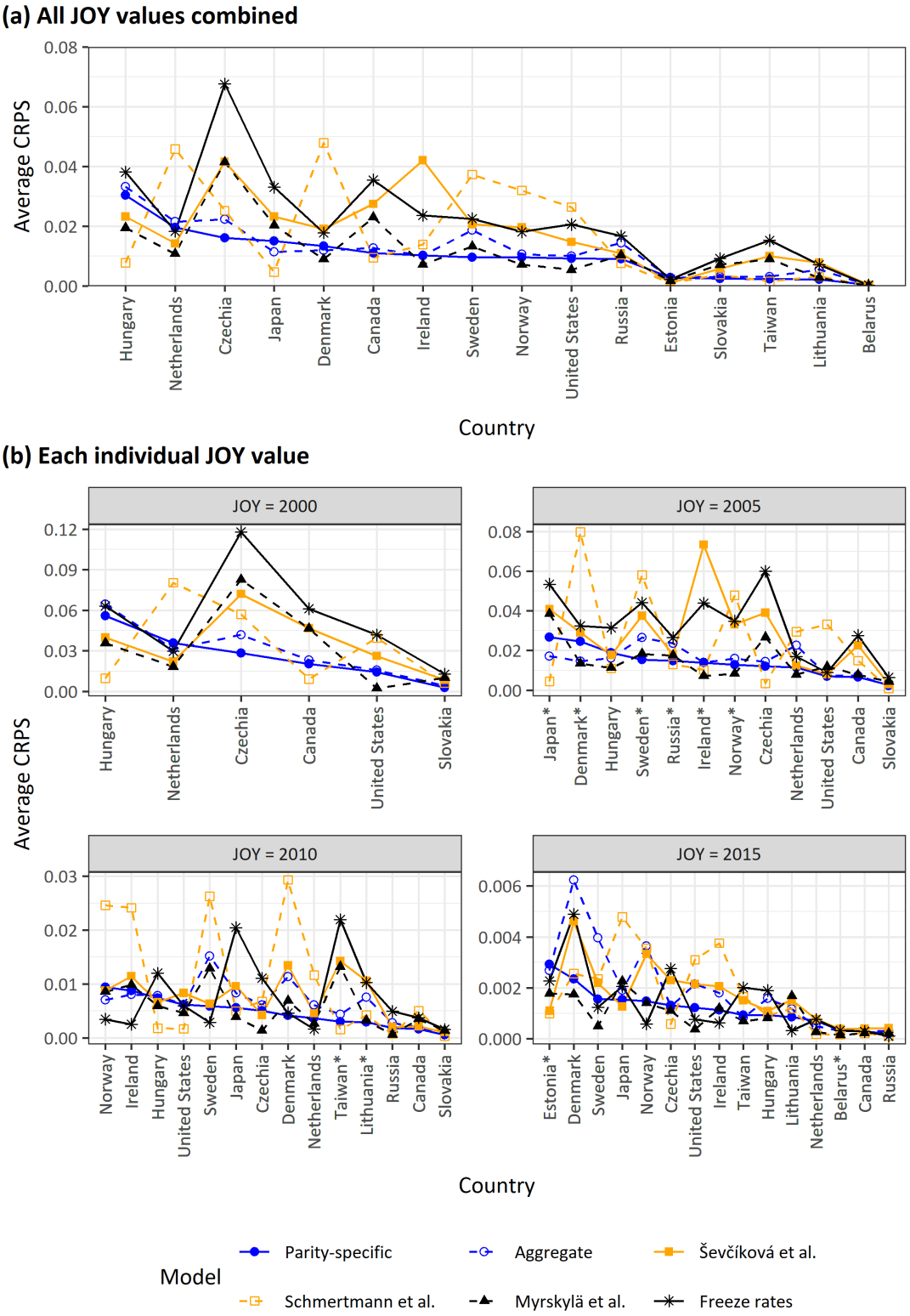

Figure 6: (a) Average CRPS against country for all *JOY* values combined, for the proposed parity-specific and aggregate models, the Ševčíková et al., Schmertmann et al. and Myrskylä et al. models, and the freeze rates approach, for the CFR forecasts. (b) Equivalent plots for each individual *JOY* value. CRPS = continuous ranked probability score; countries are ordered by decreasing CRPS according to the parity-specific model.

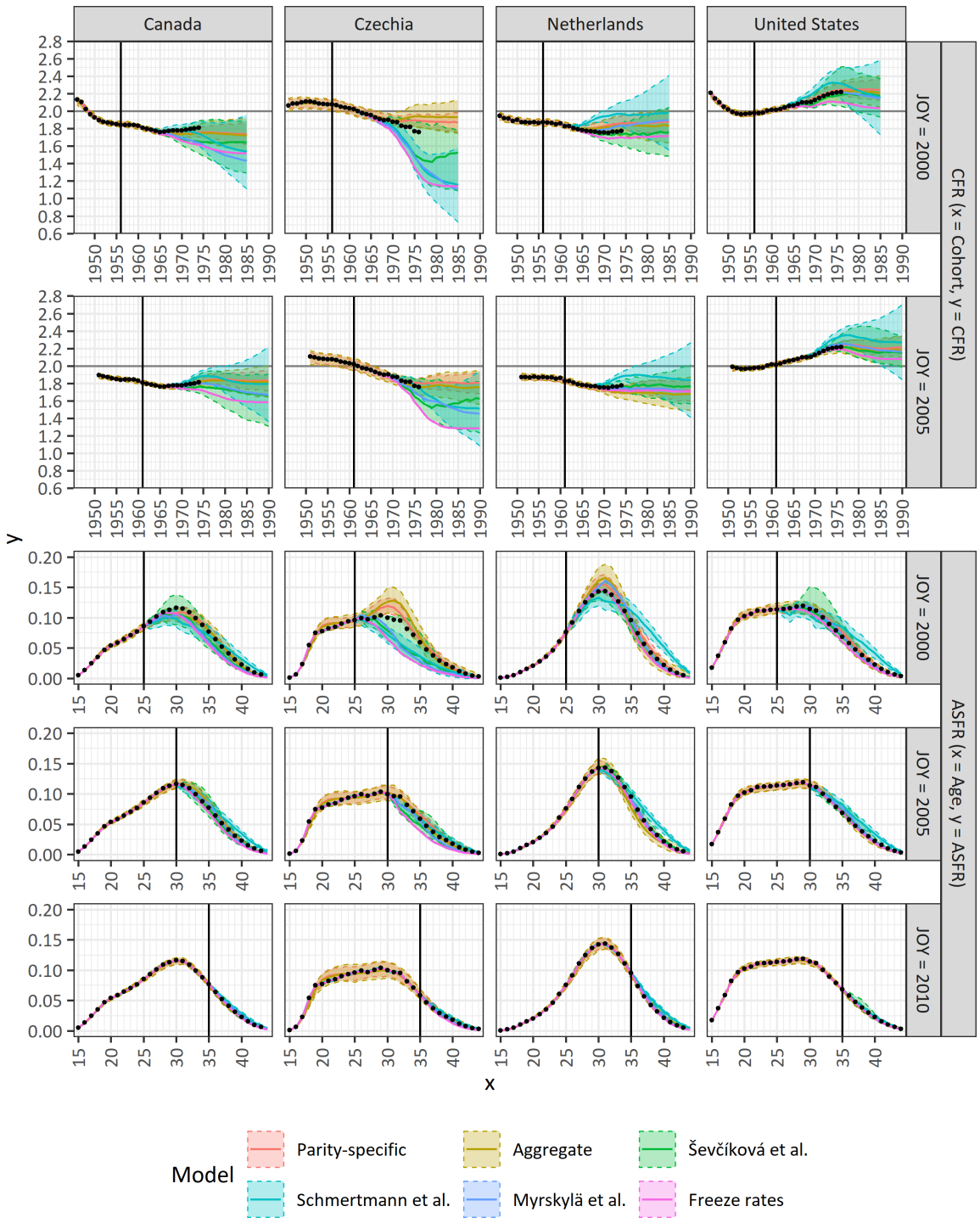

Figure 7: Rows 1 and 2: CFR posterior medians and 95% prediction intervals for the proposed parity-specific and aggregate models, the Ševčíková et al. model and the Schmertmann et al. model, and point forecasts for the Myrskylä et al. model and the freeze rates approach, for JOY values 2000 and 2005, and four selected countries; vertical lines indicate the last fully-observed cohort (born in the year $JOY - a_{max}$). Rows 3–5: For the same models and the selected countries, ASFR posterior medians and 95% prediction intervals for the 1975 cohort at JOY values 2000, 2005 and 2010; vertical lines indicate the truncation age (equal to $JOY - 1975$). Points are observed values from HFD (2024).

Appendices

A Countries included in the analysis

Table A.1: Countries included in the analysis for each JOY value

JOY	c_{min}^{JOY}	$c_{min}^{JOY^*}$	c_{max}^{JOY}	**Countries**
2000	1946	1951	1985	Canada, Czechia, Hungary, Netherlands, Slovakia, United States
2005	1951	1956	1990	Canada, Czechia, Denmark*, Hungary, Ireland*, Japan*, Netherlands, Norway*, Russia*, Slovakia, Sweden*, United States
2010	1956	1961	1995	Canada, Czechia, Denmark, Hungary, Ireland, Japan, Lithuania*, Netherlands, Norway, Russia, Slovakia, Sweden, Taiwan*, United States
2015	1961	1966	2000	Belarus*, Canada, Czechia, Denmark, Estonia*, Hungary, Ireland, Japan, Lithuania, Netherlands, Norway, Russia, Sweden, Taiwan, United States

Notes: JOY = jump-off year; c_{min}^{JOY} = minimum cohort included in original setup with $n_c = 40$ cohorts; $c_{min}^{JOY^*}$ = minimum cohort included in reduced setup with $n_c = 35$ cohorts $\left(c_{min}^{JOY^*} = c_{min}^{JOY} + 5\right)$; c_{max}^{JOY} = maximum cohort included (same for original and reduced setups). Countries are listed alphabetically, with an asterisk (*) indicating that the reduced setup is used for the particular country.

B Cumulation-based exposure approximation method

In continuation from Section 2.2, for a given cohort c, our goal is to estimate parity-specific exposures E_{ac}^i for $i \in \{0, 1, \ldots, J+\}$ and $a \in \{a_{min}, \ldots, a_{max}\}$. We define temporary exposures $\tilde{E}_{ac}^i$, exposure estimates E_{ac}, and omit the cohort subscripts for clarity. We perform the following steps as in Smallwood (2002):

1. Assume all women at age a_{min} are childless:

$$E_{a_{min}}^0 = E_{a_{min}}, E_{a_{min}}^i = 0 \text{ for } i \in \{1, 2, \ldots, J+\}.$$

2. For $a = a_{min} + 1$, obtain temporary exposures $\tilde{E}_a^i$ by updating each E_{a-1}^i according to

the following equations:

$$\tilde{E}_a^0 = E_{a-1}^0 - b_{a-1}^1;$$
$$\tilde{E}_a^i = E_{a-1}^i - b_{a-1}^{i+1} + b_{a-1}^i \text{ for } i \in \{1, 2, \ldots, (J-1)\};$$
$$\tilde{E}_a^{J+} = E_{a-1}^{J+} + b_{a-1}^J.$$

Then, scale each $\tilde{E}_a^i$ so that $\sum_i \tilde{E}_a^i = E_a$, and define this as E_a^i:

$$E_a^i = E_a \times \frac{\tilde{E}_a^i}{\sum_i \tilde{E}_a^i} \text{ for } i \in \{0, 1, \ldots, J+\}.$$

3. Repeat step 2 for $a \in \{(a_{min}+2), \ldots, a_{max}\}$.

C Comparison of parity-specific exposure estimates

In continuation from Section 2.2, we compare the exposure estimates from the cumulation-based approximation method that we use in our analysis, with the exposure estimates produced from the census- or register-based (we will refer to this as register-based for convenience) period fertility tables from HFD (2024), for all countries where this is possible. We convert the cumulation-based age-cohort estimates to age-period estimates via Lexis triangles (see Section 2.1), so that we are not introducing any additional error into the register-based estimates which are derived from period data.

In Figure C.1 we plot the age-period register-based estimates against the corresponding cumulation-based estimates for all HFD countries with available data. The correspondence is generally very good; there are a few outliers, but these are typically from countries with more unstable histories. There are consistent issues at younger ages, where the cumulation-based method tends to overestimate the exposure for parity 0 and underestimate it for higher parities, relative to the register-based method. It is important to consider that there are data quality issues to be aware of in the register-based methods too, e.g. whether the parity information was only collected from a sample of the population or everyone, the representativeness of this sample and/or the census or register as a whole, etc. Any relevant country-specific factors are summarized in the data availability spreadsheet provided in the supplementary material.

The effect on the resulting rates is also of great interest as these are the quantity that we are actually modelling – we see from Figure L.1 that for Sweden there is a small effect on the parity 0 rates but a larger impact on the rates for higher parities. However, given that these higher-order rates are very uncertain and erratic at younger ages anyway due to the small exposures and birth counts, the impact on cohort-specific model outputs such as completed family size is likely to be minimal.

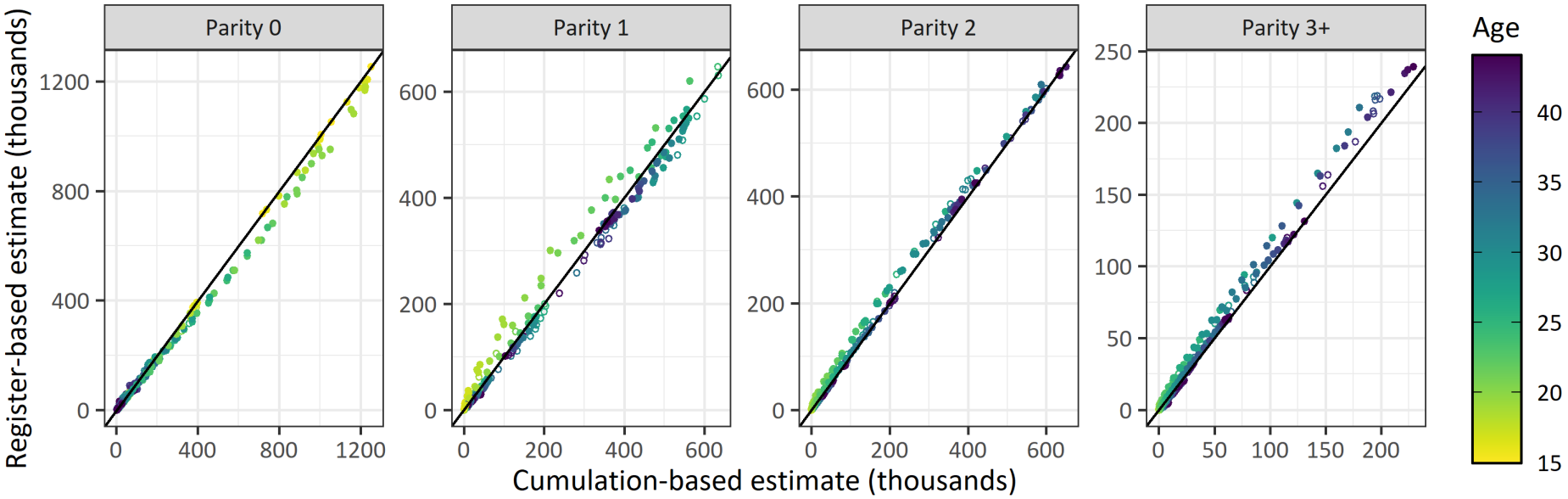

Figure C.1: Plot of parity-specific age-period census- or register-based exposure estimates (y-axis) against corresponding cumulation-based exposure estimates (x-axis) for all available countries in the HFD. The colour of the point indicates the age; open circles indicate that the cumulation-based estimate may be less reliable due to data quality issues; the scales are in thousands; the black line is where $y = x$.

D Illustration of P-spline approach for the univariate case

We illustrate the P-spline approach in Figure D.1 for the simple one-dimensional case where we are smoothing across a single variable, in this case age. Below we describe the three steps depicted in the panels of the figure:

1. Step 1 displays the cubic B-spline basis with eight basis functions.
2. Step 2 scales each of the basis functions by different weights which are constrained to change smoothly across age. Summing the values of the weighted curves for each age then gives the black smooth curve of fitted log-rates.
3. Step 3 exponentiates the fitted log-rates to give the smooth fitted rate curve.

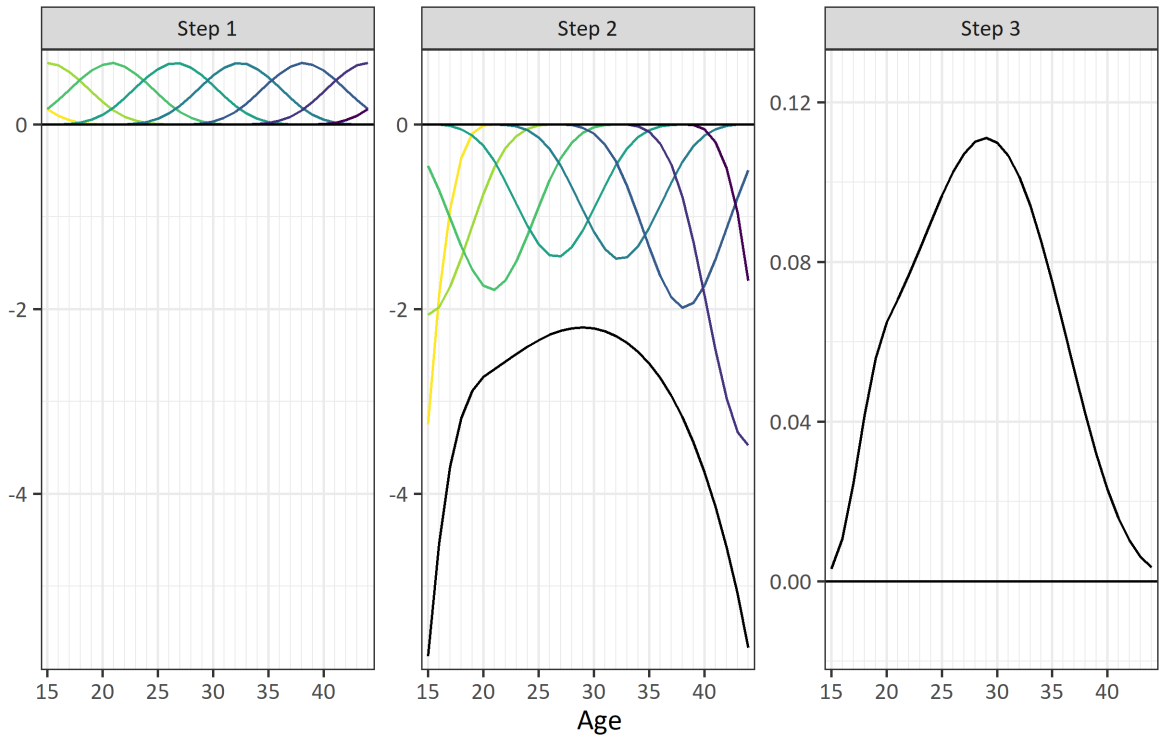

Figure D.1: Illustration of smoothing of age pattern of log-rates under a P-spline approach.

E Summary fertility measure estimates for $JOY = 2015$

Figure E.1: CFR and PPR estimates for the 1945–1971 cohorts, computed from HFD (2024). Note that only country-cohort combinations where the corresponding data is of sufficiently high quality are included.

F Prior on the proportion of women having at least one child

In continuation from Section 3.3, here we specify the second component of the Level-3 system model for the summary fertility measures. The aim of this component is to achieve plausible projections of $PPR^{0\to1}$, the proportion of women having at least one child (and equivalently $1-PPR^{0\to1}$, the proportion of women having no children), when $PPR^{0\to1}$ is close to 1 (and equivalently $1-PPR^{0\to1}$ is close to zero). To do this for a given JOY, we first extract the $PPR^{0\to1}$ estimates from the same pooled cross-country dataset from HFD (2024) used for the first component of the system model. We then estimate the kernel density of the observed $PPR^{0\to1}$ values.

As our interest is in obtaining plausible projections at the upper end of the $PPR^{0\to1}$ distribution, to determine the system model we restrict $PPR^{0\to1}$ to be greater than or equal to the value that gives rise to the last local maximum of the kernel density function (in the direction of increasing $PPR^{0\to1}$). Scaling the kernel density to take the value 1 at this maximum, we approximate the density function by a transformed Lognormal cumulative distribution function (CDF). Specifically, we identify parameters $\mu \in \{-\infty,\infty\}$ and $\sigma > 0$ such that the following quantity is minimized:

$$\sum_i \{d_i - (1 - F_X(x_i;\mu,\sigma))\},$$

where x_i is the ith $PPR^{0\to1}$ value for which the kernel density is evaluated, d_i is the kernel density estimate corresponding to x_i, and $F_X(x_i;\mu,\sigma)$ is the CDF of $X \sim \text{Lognormal}(\mu,\sigma^2)$ evaluated at x_i. We illustrate this Lognormal approximation to the scaled kernel density in Figure F.1 for $JOY = 2000$.

We then add the following quantity to the log-likelihood in the model, where the sum is over all cohorts c:

$$\sum_c \{\log(1 - F_X(PPR_c^{0\to1};\mu,\sigma) + 0.00001)\}.$$

In this way, large values of $PPR^{0\to1}$ are penalized with increasing strength as $PPR^{0\to1}$ increases, according to the Lognormal CDF approximation to the scaled density of historical $PPR^{0\to1}$ estimates. Where the CDF takes a value of 0, i.e. approximately where $PPR^{0\to1}$ takes a value below that which gives rise to the last local maximum of the density function, there is essentially no effect on the log-likelihood (as $\log(1.00001) \approx 0$). But as $PPR^{0\to1}$ increases from this value (and the value of the CDF approaches 1), an increasingly negative quantity is added to the log-likelihood, corresponding to a stronger penalization. Where the CDF takes a value of 1, i.e. where $PPR^{0\to1}$ is very close to 1, the effect on the log-likelihood is at its most negative at $\log(0.00001) \approx -11.5$, indicating that almost universal childbearing is highly unlikely. The addition of 0.00001 ensures that the quantity added to the log-likelihood is always finite, thus avoiding the computational issues associated with adding a value of $\log(0) = -\infty$ while having a similar impact.

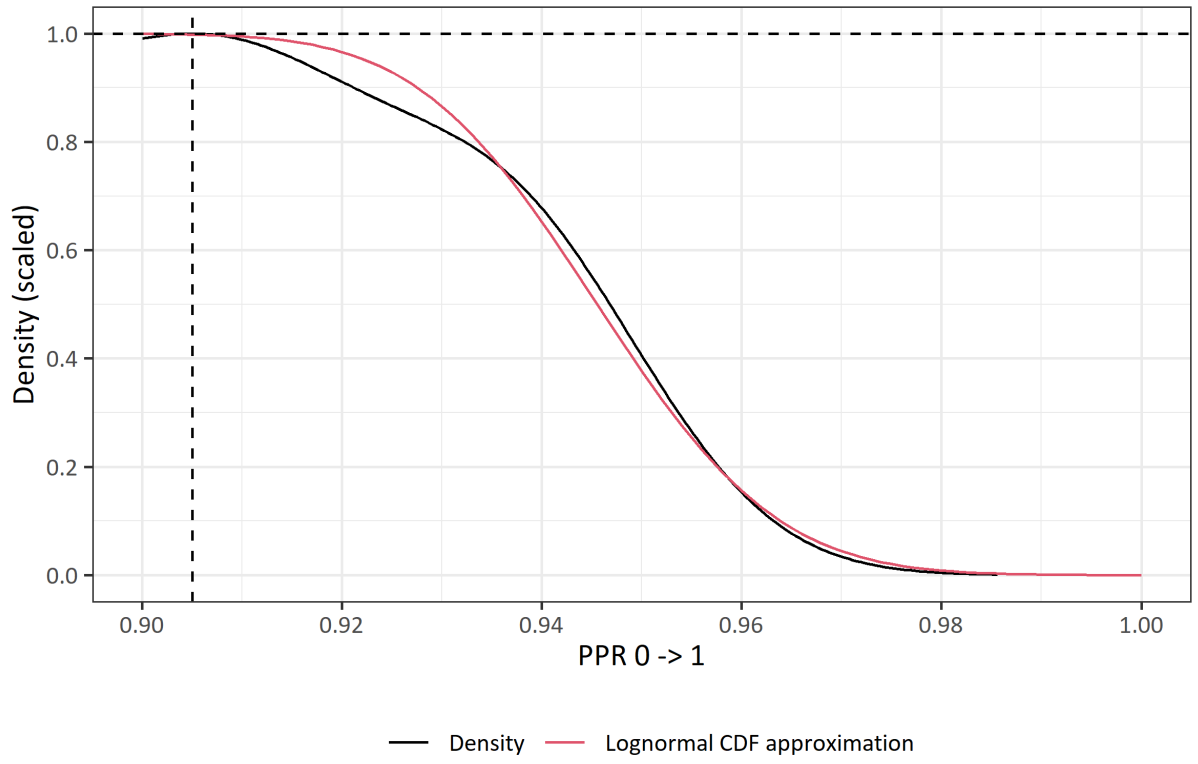

Figure F.1: Illustration of the $PPR^{0\to1}$ prior for $JOY = 2000$. The density (black line) is scaled to have a value of 1 at the last local maximum (in the direction of increasing $PPR^{0\to1}$); the value at which this maximum is achieved is indicated by the vertical dashed line, while the horizontal dashed line is at $y = 1$.

G Fitting the Schmertmann et al. model

We fit the model of Schmertmann et al. (2014a) by adapting the code downloaded from the project website (Schmertmann et al. 2014b). The main changes we make are in relation to the data used in model fitting. Firstly, for consistency with the data used to fit our models, we use the *asfrVH* (fertility rates by birth cohort and age) and *exposVH* (female population exposure by birth cohort and age) format from the HFD instead of the *asfrVV*/*exposVV* format (rates/exposures by calendar year, age and birth cohort) used in the original code. We also use the latest HFD data and therefore include additional countries that have since been added to the database (and remove them from the additional non-HFD data provided in the original code if they now have HFD data available).

As we have included countries in the "reduced" setup of the model which fits to 35 cohorts instead of 40 (see Section 2.1), we create an adjusted version of the Schmertmann et al. model which fits to the smaller number of cohorts. We also create a second adjusted version of the model which ignores past uncertainty about the ASFRs, thus treating them as known. This is required for a fair comparison in terms of summary statistics (see Section 4.2.1). In this way, we generate a maximum of four forecast outputs per JOY: one set for those countries with the original 40-cohort setup, containing two forecast outputs which either incorporate or ignore past ASFR uncertainty; and an equivalent set for those countries with the reduced 35-cohort setup, if any are present in the particular JOY.

H Fitting the Ševčíková et al. model

As the United Nations population projections take a period perspective, the adjustments required to incorporate the model of Ševčíková et al. (2016) into our comparative analysis are more complex than those for the Schmertmann et al. (2014a) model. We use fertility rate estimates by calendar year and age from the HFD, i.e. the *asfrRR* format. For a given *JOY*, using the `bayesTFR R` package (Ševčíková et al. 2011) which implements the hierarchical Bayesian TFR projection model of Alkema et al. (2011), we first generate 60,000 posterior parameter samples from the phase 2 and phase 3 models, fitting the model to the past 30 years of annual TFR data (computed from the HFD ASFR estimates); these specifications were found to be optimal in Bohk-Ewald et al. (2018b). Discarding the first half of these parameter samples as burn-in, we then generate 30,000 posterior TFR samples for each country for the next 30 years (again, annually), to be able to complete all cohorts. We then thin these samples to retain 500 posterior TFR samples for each country and forecast year.

Next we deconstruct the posterior TFR samples into ASFR samples using the `bayesPop R` package (Ševčíková and Raftery 2016) which implements the method of Ševčíková et al. (2016). We specify the global pattern as the variant that was found to be optimal in Bohk-Ewald et al. (2018b), and then generate samples of the percent ASFR (PASFR) by single year of age which we then apply to disaggregate the TFR samples. Following this it is necessary to appropriately convert the ASFR samples from age-period to age-cohort rates. We do this by first multiplying the ASFR samples by the corresponding age-period exposures from the HFD (*exposRR* format) and the WPP (United Nations, Department of Economic and Social Affairs, Population Division 2024) to get age-period birth samples. We then convert these to approximate age-cohort birth samples via Lexis triangles (see Section 2.1). Dividing these by the age-cohort exposures which we use to fit our proposed model then gives age-cohort ASFR samples to use in our comparative analysis.

There are three important caveats to bear in mind regarding our implementation of the model of Ševčíková et al. (2016):

1. The period to cohort conversion is only approximate, so where it is less accurate the model may perform more poorly compared to an assessment of predictive accuracy in the original period context.

2. In our cohort setup (Section 2.1) we take observed age-cohort combinations to be where $a + c \leq JOY$ (age = a, cohort = c). However in a period setup, the data relating to the case when $a + c = JOY$ actually straddles the years JOY and $JOY + 1$. So the exact jump-off point has a different interpretation in the period and cohort contexts. For consistency with the cohort setup, we treat the case when $a + c = JOY$ as being observed perfectly in the period setup, even though it is in fact partly forecasted.

3. We have not accounted for uncertainty about the historical TFR observations and so are assuming that the ASFRs before jump-off (where $a + c \leq JOY$) are perfectly observed. This means that our calculated coverage in Table 1 is likely to be a slight underestimate (Liu and Raftery 2020).

I Additional rate forecasts for $JOY = 2015$

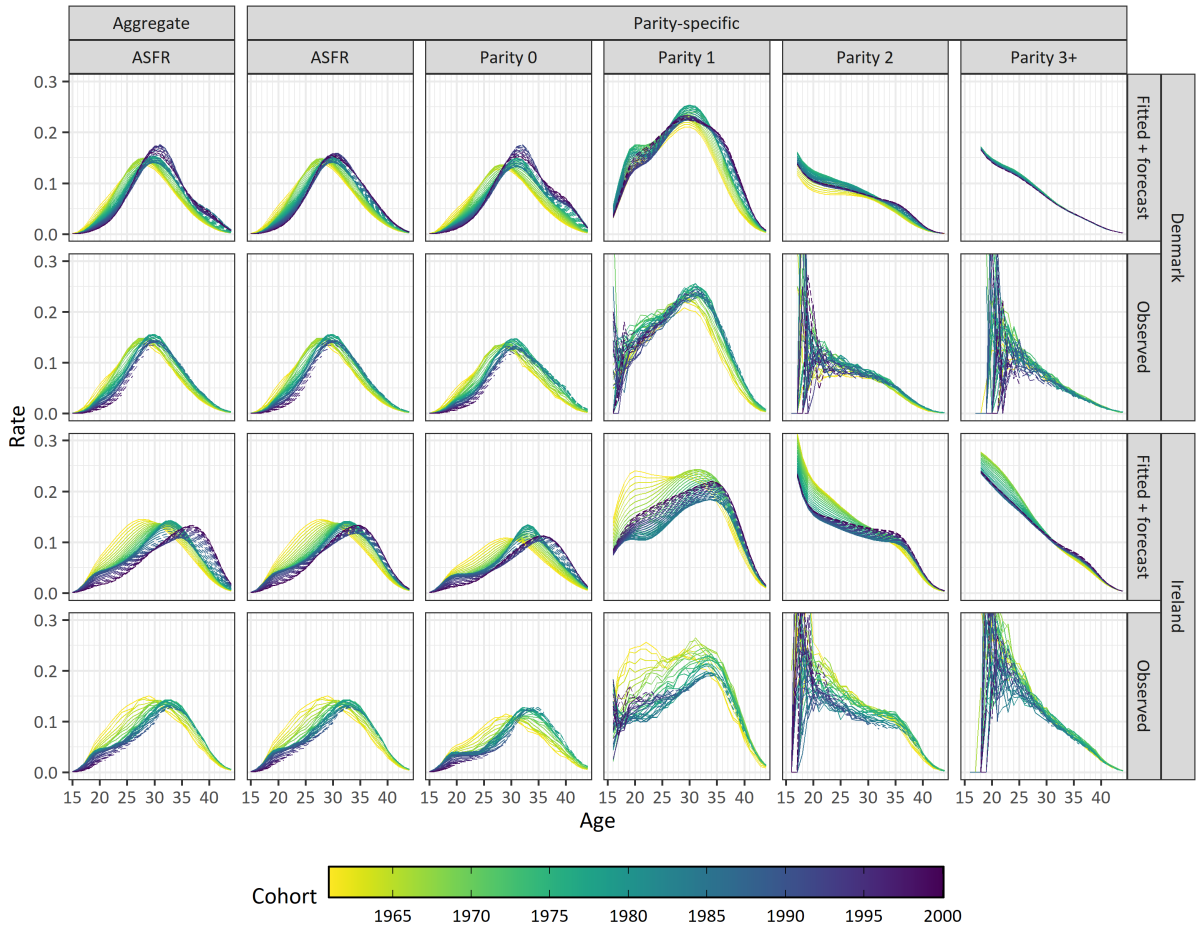

Figure I.1: Row 1: Denmark posterior median ASFRs and parity-specific rates from the aggregate and parity-specific fertility projection models for $JOY = 2015$; Row 2: corresponding observed rates from HFD (2024). Rows 3 and 4: Equivalent plots for Ireland. Dashed lines indicate values after the year JOY.

J Sensitivity analysis: Prior on σ_{SM}

In continuation from Section 3.3, the current prior for a given true summary measure SM, denoted by θ_c^{SM} for cohort c, is a normal prior for a first-order random walk:

$$\theta_c^{SM} \sim N\left(\theta_{c-1}^{SM}, \sigma_{SM}^2\right), c = \left(c_{min}^{JOY} + 1\right), \ldots, c_{max}^{JOY}.$$

This prior fixes the standard deviation (SD) σ_{SM} for a given JOY value, which is estimated using historical HFD data. In this appendix we investigate the sensitivity of the forecasts to this prior by including σ_{SM} as a parameter in the model, therefore obtaining posterior samples of the SD rather than fixing it beforehand.

We specify a non-informative uniform prior for σ_{SM}, with bounds given by the lowest and highest country-specific historical standard deviations for a given JOY[1]. To assess the effect of this uniform prior on our results, for each JOY we fit our proposed parity-specific and aggregate models to three countries to cover a range of contexts: the one with the highest SD and the one with the lowest SD (where $SM = CFR$), and the one with the closest to average SD (again where $SM = CFR$, i.e. the current fixed value of σ_{CFR})[2]. We demonstrate this for $JOY = 2015$ in Figure J.1, which plots the country-specific SD values for each summary measure and overlays the corresponding current fixed SD values. From the CFR panel, it is therefore clear that the USA (highest), Norway (lowest) and Estonia (closest to average) are selected for fitting. We specify the uniform prior for σ_{SM} as:

$$\sigma_{SM} \sim U\left(\hat{\sigma}_{SM}^l, \hat{\sigma}_{SM}^u\right),$$

where $\hat{\sigma}_{SM}^l$ and $\hat{\sigma}_{SM}^u$ are the lowest and highest country-specific SD values respectively. From Figure J.1 we can see for example that for the CFR these are the values corresponding to Norway and the USA, and for $PPR^{0\to1}$ these are the values for Sweden and Belarus.

[1] We require the country to have at least six consecutively observed cohorts up to and including the year JOY, so that we could in theory fit our model to the country.

[2] We choose these three countries from the set of those that we are already fitting to for the particular JOY value and have therefore ascertained that the required data is available.

We fit these alternative versions of the parity-specific and aggregate models, i.e. incorporating the uniform prior on σ_{SM} for each available summary measure, to the 12 country-JOY combinations. We first assess the posterior distributions of the σ_{SM} parameters by plotting the resulting densities in Figure J.2. We see that across all combinations, despite the non-informative nature of the prior, the σ_{SM} density is strongly positively skewed, pushing towards the lower bound of the uniform prior. Consequently, the introduction of the prior leads to a smaller estimate of each σ_{SM} parameter compared to the original implementation of the model (the fixed SD values are indicated by the dashed lines). This in turn leads to a tighter first-order random walk prior on the summary measures and therefore narrower prediction intervals (PIs). Indeed, the overall summary statistics for the CFR forecasts (not shown) indicate that the coverage of the 50% and 90% PIs is slightly reduced (tending to become closer to the respective nominal levels as the original coverage is above nominal), while the other measures of predictive accuracy all worsen.

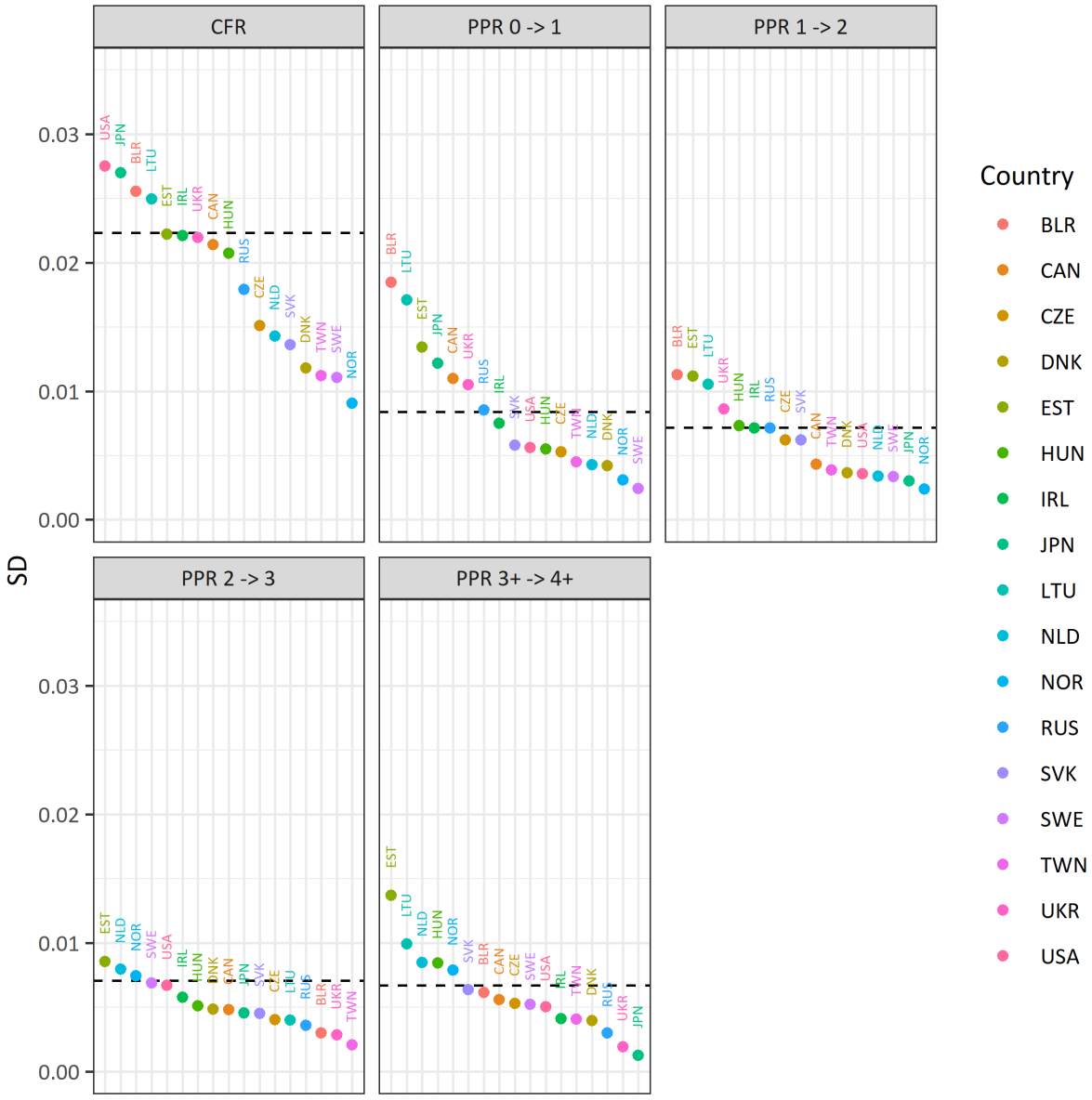

Figure J.1: Country-specific SD values for JOY = 2015, with SD based on all countries combined ('average SD', used in original model) indicated by horizontal dashed lines.

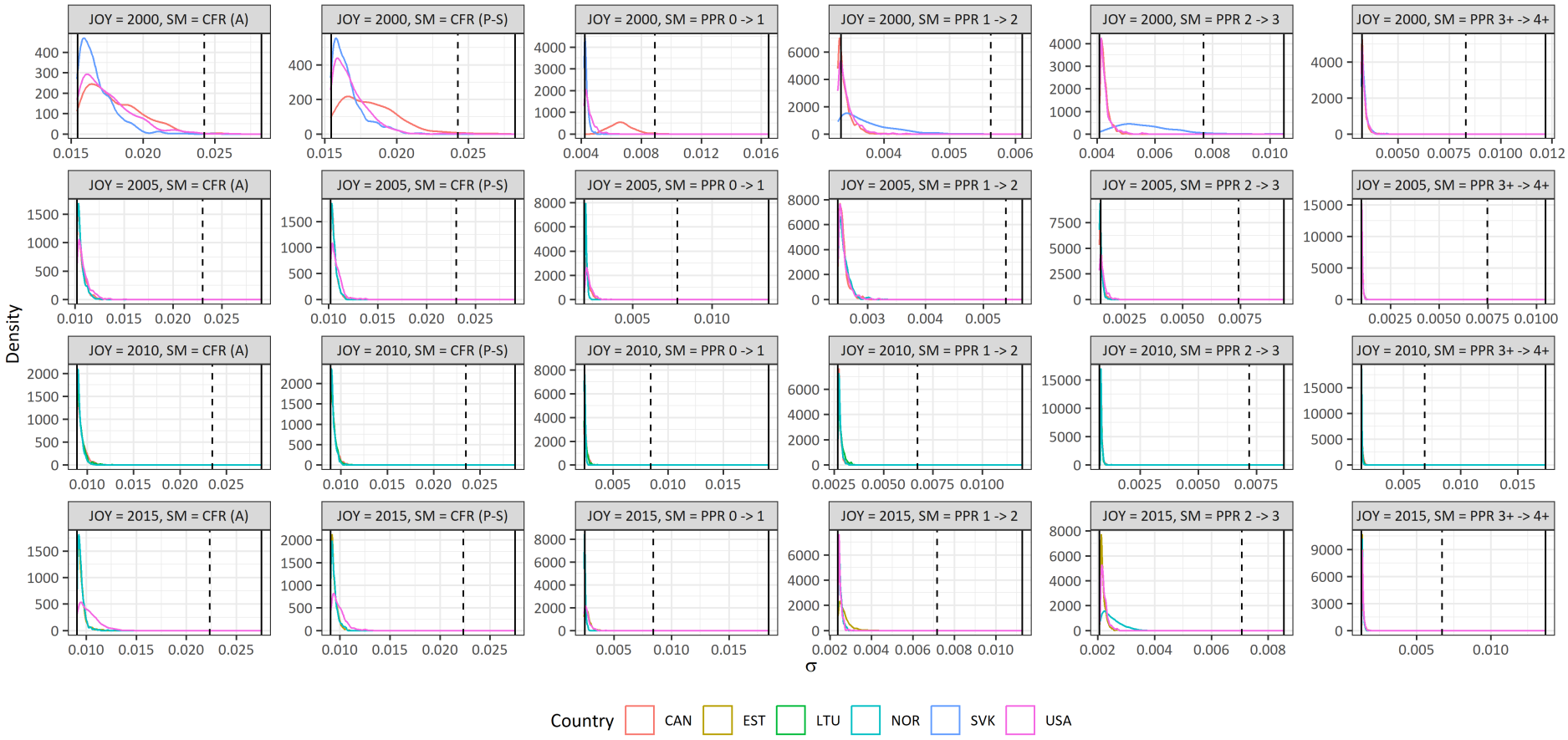

Figure J.2: Density plots of σ_{SM} for each JOY-SM combination (panels, where each row is a different JOY and each column is a different SM). In each panel, the coloured lines correspond to the different countries fitted to for the particular JOY, the vertical lines indicate the lower and upper bounds of the uniform prior ($\hat{\sigma}^{l}_{SM}$ and $\hat{\sigma}^{u}_{SM}$), and the dashed vertical line indicates the 'average SD' used in the original model; A = aggregate, P-S = parity-specific.

To illustrate this visually, in Figure J.3 we plot the coverage of the 50% and 90% PIs, and the average continuous ranked probability score (CRPS; see Section 4.2.1), for the CFR forecasts from the original and alternative versions of the aggregate and parity-specific models. For the coverage of the 50% PI in particular (row 1), in the vast majority of cases there is a movement towards the nominal value (dashed line), while the coverage of the 90% PI (row 2) often does not change or moves further away from the nominal level). However, the average CRPS, which takes into account the entire predictive distribution of the CFR as opposed to simply its lower and upper quantiles, indicates that predictive accuracy tends to stay roughly the same or worsen, improving substantially in only a few cases. The analogous plot for the ASFR forecasts is given in Figure J.4. In contrast with Figure J.3, the coverage of the 50% and 90% PIs is already below nominal level for the original model, moving further away from nominal in most cases. The CRPS improves in roughly half of the cases, and stays the same or worsens in the rest. Therefore there is no consistent improvement in the measures across countries and model type.

In conclusion, our results appear to be reasonably robust to the addition of a prior on σ_{SM}. Although there are improvements in coverage at the 50% level for the CFR forecasts, this is not seen at the 90% level and is often at the cost of poorer predictive accuracy measured by the CRPS, with findings worsening for the ASFR forecasts. Extensions to this sensitivity analysis could include testing other prior specifications for σ_{SM}; however, the clear desire that the parameter has to be small (see Figure J.2) could cause convergence issues if an unbounded prior were specified, as this would give the parameter the freedom to be even closer to zero.

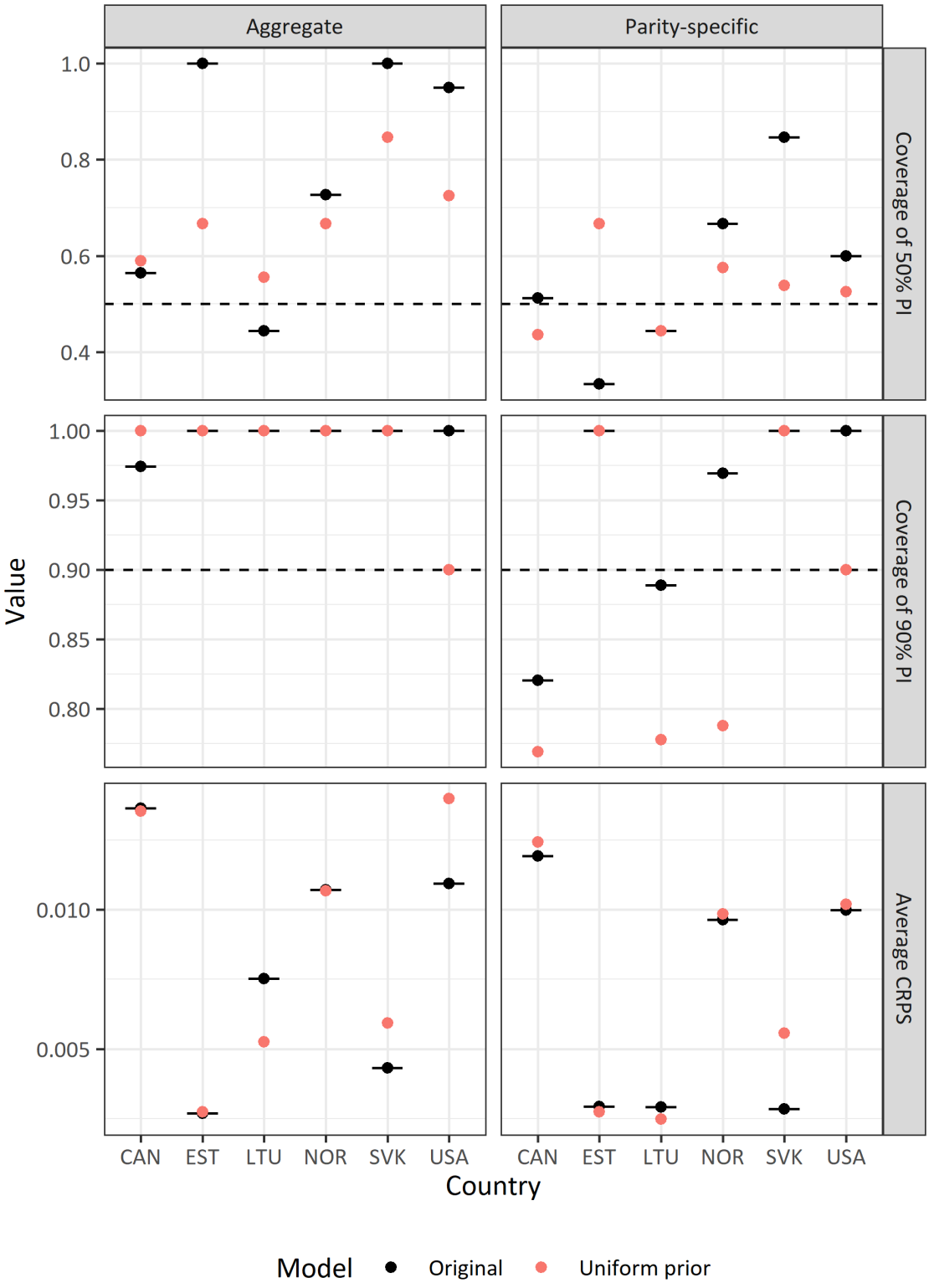

Figure J.3: Plot of coverage of 50% and 90% PIs, and average CRPS (rows), against country, for the CFR forecasts from the aggregate and parity-specific models (columns) under the original and uniform priors. The horizontal dashed lines indicate the nominal coverage values. PI = prediction interval, CRPS = continuous ranked probability score.

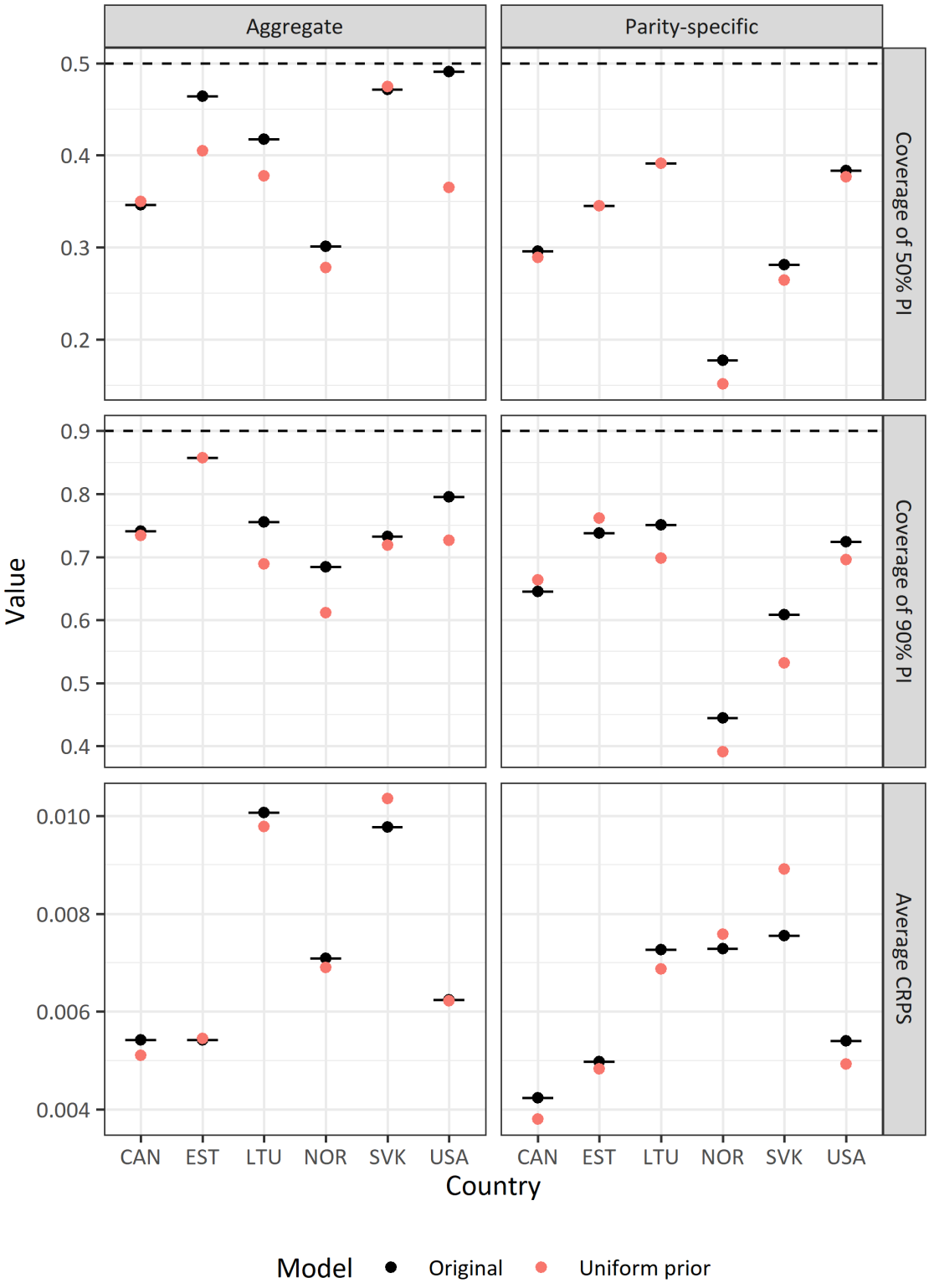

Figure J.4: Plot of coverage of 50% and 90% PIs, and average CRPS (rows), against country, for the ASFR forecasts from the aggregate and parity-specific models (columns) under the original and uniform priors. The horizontal dashed lines indicate the nominal coverage values. PI = prediction interval, CRPS = continuous ranked probability score.

K Sensitivity analysis: Knot spacing

In this appendix we investigate the sensitivity of the forecasts to the spacing of the knots in the age and cohort basis functions. In continuation from Section 3.2.1, the original implementation of the model spaces the knots every five years in the age and cohort basis functions, thereby having basis dimensions of $P = 8$ (for the age basis) and $Q = 10$ (for the cohort basis). We test the effect of spacing the knots more or less frequently than this, considering four additional combinations which space the knots roughly every 3 or 7.5 years in one or both of the age and cohort dimensions. The combinations that we consider are summarized below in Table K.1:

P	Q	**Description**
8	10	Original knot spacing (knots spaced every 5 years in both dimensions)
12	15	Knots spaced roughly every 3 years in both dimensions
8	15	Original knot spacing in age dimension Knots spaced roughly every 3 years in cohort dimension
12	10	Knots spaced every 3 years in age dimension Original knot spacing in cohort dimension
6	7	Knots spaced roughly every 7.5 years in both dimensions

Table K.1: Summary of the basis dimension specifications that we consider, together with their corresponding description in terms of knot spacing specifications. P = number of basis functions in age dimension; Q = number of basis functions in cohort dimension.

We test these alternative basis dimension specifications by fitting the appropriately adjusted parity-specific and aggregate models to the six countries included for $JOY = 2000$ (see Table A.1), namely Canada, Czechia, Hungary, the Netherlands, Slovakia and the United States. We choose a single JOY value because the presence of four alternatives means that it would be highly computationally intensive to include all country-JOY combinations; given this, selecting the earliest JOY value provides the largest amount of holdout data to assess forecast performance. We note that the most complex option ($P = 12$, $Q = 15$) did not converge for Canada, Czechia and the Netherlands, while the $P = 8$, $Q = 15$ option did not converge for Hungary and Czechia; the results for these combinations are therefore omitted.

In Figure K.1 we plot the coverages of the 50% and 90% prediction intervals (PIs) and the

average continuous ranked probability score (CRPS) (rows) for the CFR forecasts under the aggregate and parity-specific models (columns), for each country and knot spacing specification; Figure K.2 presents the analogous plot for the ASFR forecasts. We observe that for the CFR forecasts none of the alternative basis dimension combinations consistently improve coverage and the CRPS compared to the original specification, and for the ASFR forecasts the performance is often worse than the original (this is similar to the findings from Appendix J). It is important to have confidence that the model will converge, which is not possible for the two options with $Q = 15$; the remaining options can improve or worsen predictive performance depending on the country and model type, and in some cases the coverage may improve but the CRPS worsen, and vice versa.

In conclusion, of the alternative knot spacing specifications under consideration in this sensitivity analysis, there is no option that provides a consistent improvement in coverage and/or predictive accuracy. The forecast performance can get better or worse, seemingly dependent on the stability of past rates. Therefore, although alternative basis dimension specifications may improve the results for an individual country-JOY-model combination, there is no evidence to support one specification being preferable in general.

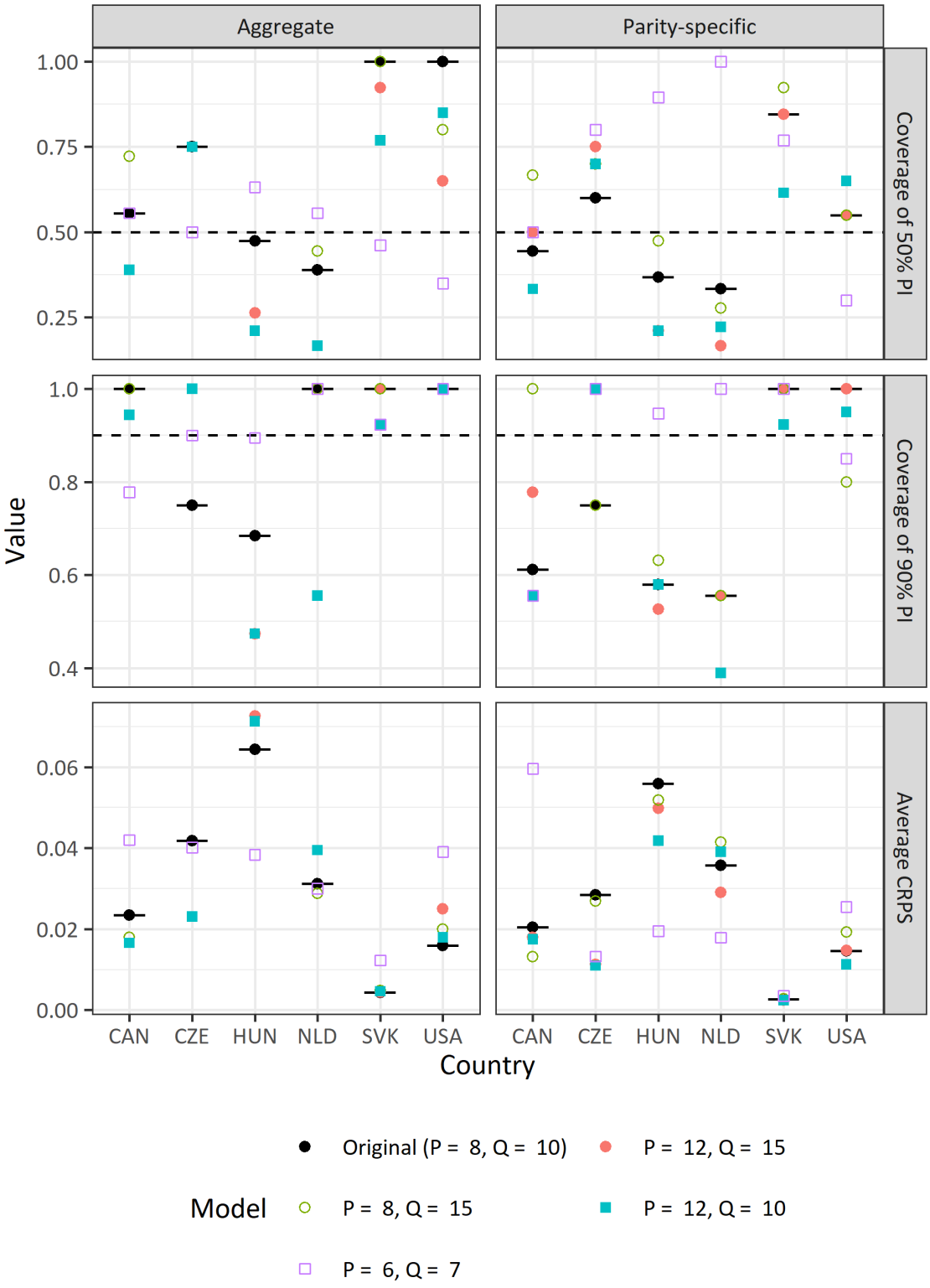

Figure K.1: Plot of coverage of 50% and 90% PIs, and average CRPS (rows), against country, for the CFR forecasts for $JOY = 2000$ from the aggregate and parity-specific models (columns) under different knot spacing specifications. The horizontal dashed lines indicate the nominal coverage values. PI = prediction interval, CRPS = continuous ranked probability score, NbA (NbC) = number of basis functions in age (cohort) dimension.

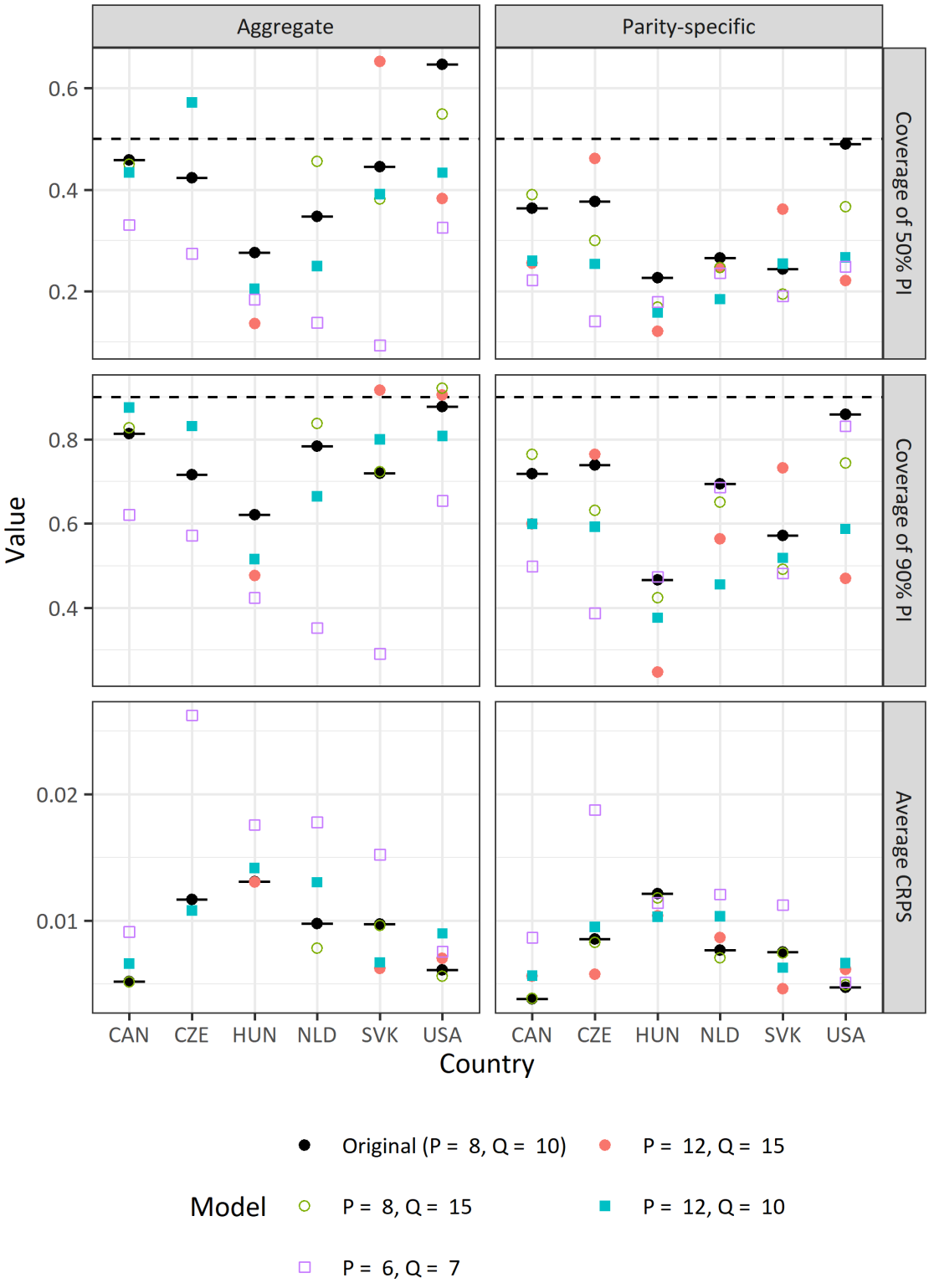

Figure K.2: Plot of coverage of 50% and 90% PIs, and average CRPS (rows), against country, for the ASFR forecasts for $JOY = 2000$ from the aggregate and parity-specific models (columns) under different knot spacing specifications. The horizontal dashed lines indicate the nominal coverage values. PI = prediction interval, CRPS = continuous ranked probability score, NbA (NbC) = number of basis functions in age (cohort) dimension.

L Sensitivity analysis: Exposure approximation

In this appendix we investigate the impact of fitting the parity-specific model using the observed census- or register-based (we will refer to this as register-based for convenience) exposure estimates from the HFD instead of the cumulation-based exposure approximations that we use in our original model (see Section 2.2). As explained in Section 2.2, there are very few countries in the HFD with a sufficiently long and detailed time series of direct parity-specific exposure estimates. In fact, only Hungary and Sweden have the data required to fit our proposed model for any of the four JOY values, and even then it is not possible to include all JOY values but just 2005 (reduced setup), and 2010 and 2015 (original setup). We first convert the annual age- and parity-specific register-based exposure estimates to approximate parity-specific age-cohort exposures via Lexis triangles (see Section 2.1), and then rescale them so that the age-cohort totals match the values in our original model for consistency.

It is then possible to compute the conditional parity-specific age-cohort fertility rate estimates calculated using these register-based exposures, and compare them with the equivalent rates calculated using the cumulation-based exposures. We do this for the Swedish rates used to fit the model for $JOY = 2015$, presenting the visual comparison in Figure L.1. We see that the parity 0 rates appear practically identical, while the higher-order rates are noticeably lower and considerably less erratic at younger ages. These observations are consistent with our findings in Appendix C where we compared the exposure estimates produced by the two methods, finding that the register-based exposures were slightly smaller for parity 0, and slightly larger for parities 1 and above, at the youngest ages.

It is necessary to adjust our parity-specific model slightly so that we incorporate the register-based exposure estimates as the known parity-specific exposures for the observed age-cohort combinations, but still use the cumulation-based method to approximate the exposures for the future age-cohort combinations. We consider two variants: Variant 1, which retains the original informative prior on the parity-specific dispersion parameters (see Section 3.2.2); and Variant 2, which applies the vague prior used for the aggregate

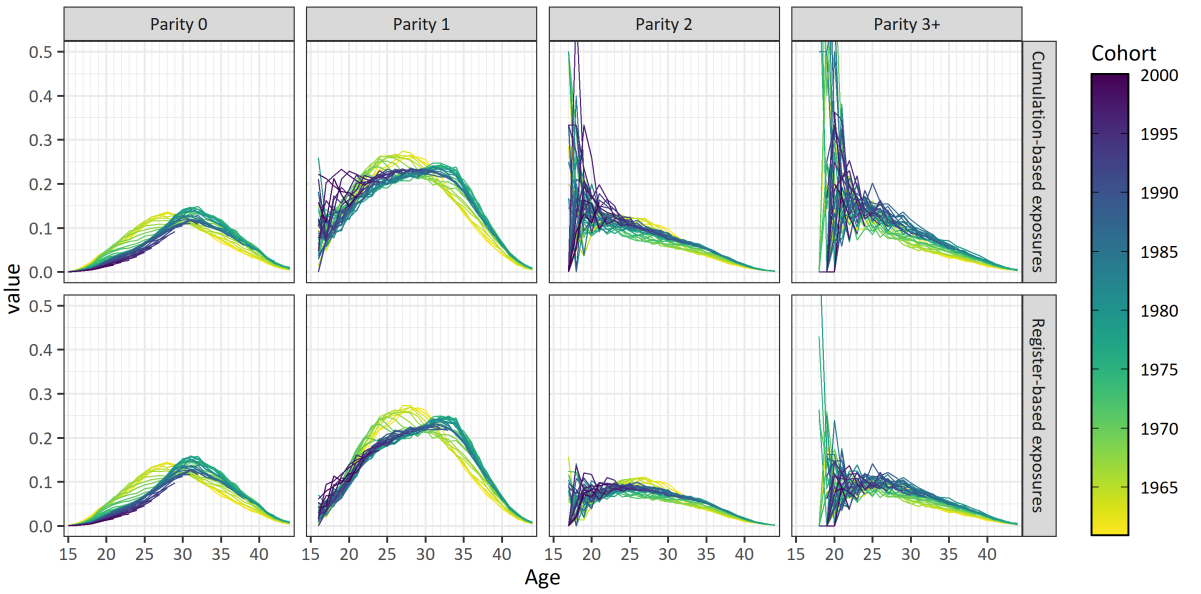

Figure L.1: Sweden parity-specific rate estimates under cumulation-based exposures (top row) and census- or register-based exposures (bottom row), 1961–2000 cohorts, used to fit the model for $JOY = 2015$.

dispersion parameter to the parity-specific dispersion parameters. During the model fitting process we identified issues with both variants in that convergence was not always achieved on the first model run. We also observed that the forecasts of the parity-specific summary measures sometimes displayed a lack of smoothness, due to treating the observed register-based parity-specific exposures as known rather than generated from the smooth rates under the cumulation-based method.

Figure L.2 presents the coverage of the 50% and 90% prediction intervals (PIs) and the average continuous ranked probability score (CRPS) for the CFR forecasts under the original model, Variant 1 and Variant 2, for the three JOY values in the columns, and the two countries. We see that the results are quite mixed: Variant 1 tends to give almost identical or slightly worse coverages for both levels, but sometimes provides an improvement in the average CRPS; Variant 2 again often has the same coverage, but does provide some large improvements such as for $JOY = 2005$ (Sweden and Hungary at the 90% level, just Sweden at the 50% level) and $JOY = 2010$ (just Sweden at the 50% level). However, the CRPS for Variant 2 is often worse, apart from a considerable improvement for Hungary when $JOY = 2005$. Therefore, as we have seen in the earlier sensitivity analyses, there is a lack of consistent

improvement across countries, JOY values and metrics. The analogous plot for the ASFR forecasts is shown in Figure L.3. The coverages tend to be similar for Variant 1 and worse for Variant 2, and are again all below the nominal level as we have seen in Appendices J and K. The CRPS shows an improvement for both variants for $JOY = 2005$, but Variant 2 performs much worse than the original method for the other JOY values.

This sensitivity analysis appears to indicate that using register-based exposures rather than cumulation-based exposures has a generally negative impact on predictive performance. We would actually expect there to be an improvement, as it should be preferable to use exposure estimates that we believe to be more accurate. However, with our limited testing data we find that Variant 1 has little impact, and Variant 2, which expresses a high level of confidence in the register-based exposures, performs very poorly. Furthermore, both variants present convergence issues and unsmooth forecasts. Possibly the prior on the parity-specific dispersion parameters should express a moderate level of confidence in the register-based exposures, somewhere in-between Variants 1 and 2. Additionally, to improve convergence, it may be the case that the register-based exposures should only appear in the likelihood statement and not when generating the summary measures within the model; this would avoid the change in methodology when we move from the observed to the future age-cohort combinations, and would also resolve the issue of the unsmooth forecasts. In conclusion, further investigation is clearly needed to identify the optimal way to incorporate the register-based exposures into our proposed model, and to expand the pool of testing data as much as possible to enable firmer conclusions to be drawn.

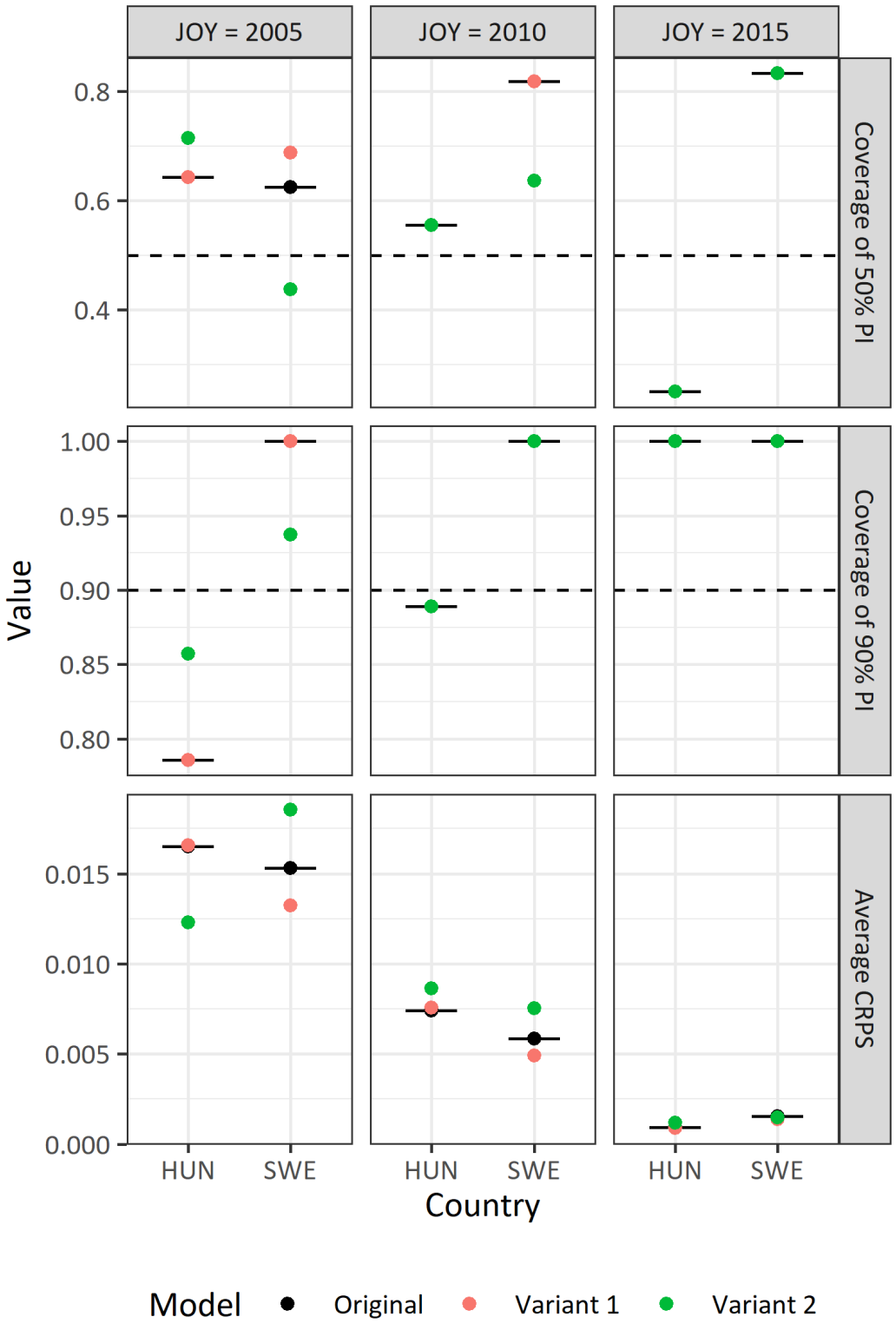

Figure L.2: Plot of coverage of 50% and 90% PIs, and average CRPS (rows), against country, for the CFR forecasts from the parity-specific model for *JOY* values 2005, 2010 and 2015 (columns) under the original and alternative exposure specifications. The horizontal dashed lines indicate the nominal coverage values. PI = prediction interval, CRPS = continuous ranked probability score.

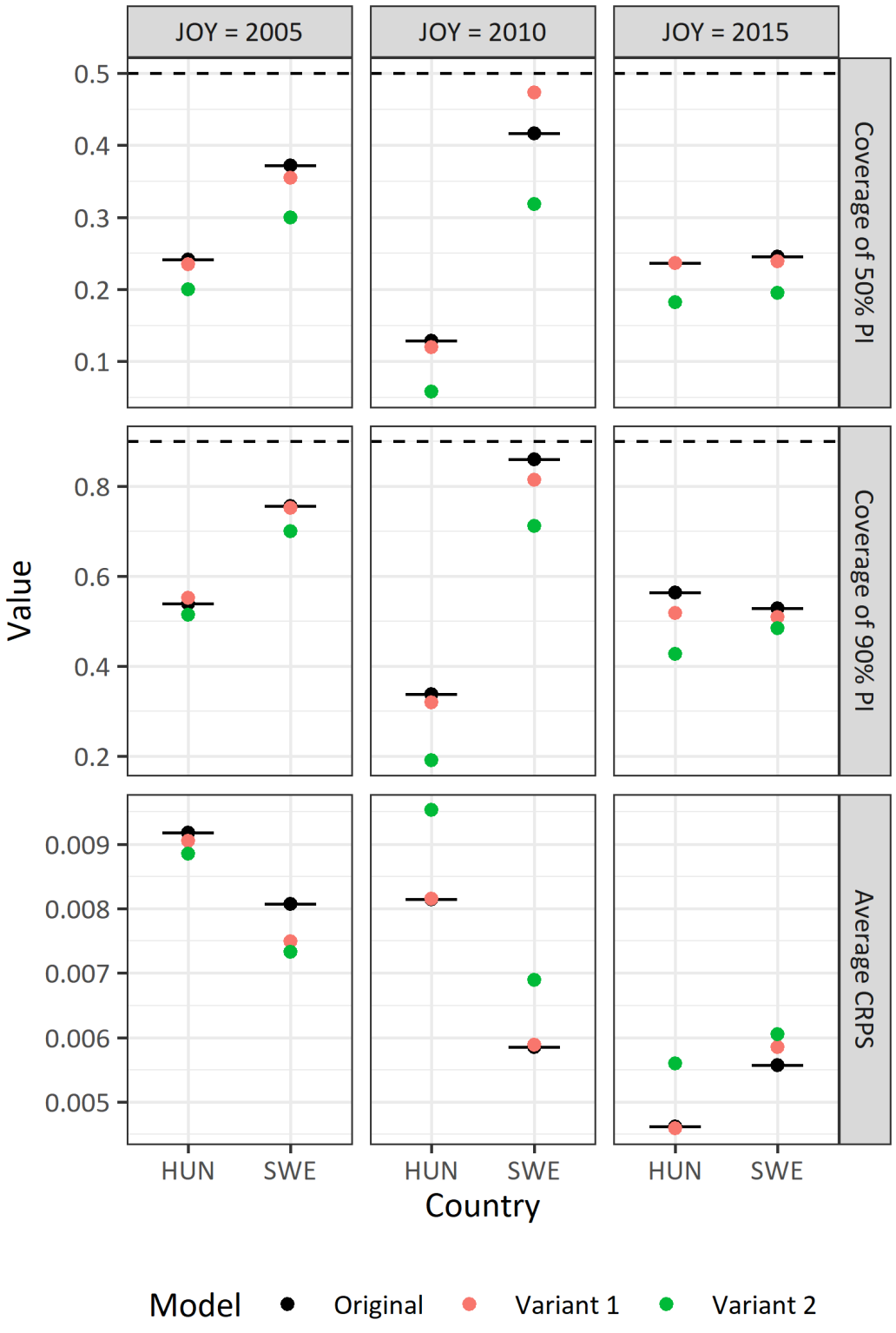

Figure L.3: Plot of coverage of 50% and 90% PIs, and average CRPS (rows), against country, for the ASFR forecasts from the parity-specific model for *JOY* values 2005, 2010 and 2015 (columns) under the original and alternative exposure specifications. The horizontal dashed lines indicate the nominal coverage values. PI = prediction interval, CRPS = continuous ranked probability score.